DOCUMENT RESUME

ED 209 810 EC 140 324

AUTHOR Reschly, Daniel J.
TITLE Nonbiased Assessment.
INSTITUTION Iowa State Dept. of Public Instruction, Des Moines.; Iowa State Univ. of Science and Technology, Ames. Dept. of Psychology.
PUB DATE Dec 80
NOTE 83p.

EDRS PRICE MF01/PC04 Plus Postage.
DESCRIPTORS Compliance (Legal); Court Litigation; *Disabilities; Elementary Secondary Education; Evaluation Methods; Federal Legislation; Mild Mental Retardation; *Minority Groups; *Sociocultural Patterns; *Test Bias; Test Interpretation
IDENTIFIERS System of Multicultural Pluralistic Assessment; Wechsler Intelligence Scale for Children (Revised)

ABSTRACT

Issues surrounding the concept of nonbiased assessment, particularly with economically disadvantaged minority students, are analyzed. The legal requirements are focused on thorough discussion of litigation and legislation. The second section explores research on the use of the Wechsler Intelligence Scale for Children-Revised (WISC-R). Research touches on the many definitions of bias, construct validity/content bias, item bias, atmosphere bias, and bias in test use. The author states that conclusions regarding the validity and bias of the WISC-R with minorities vary depending on the definition of bias. A definition is proposed based on the outcomes of test use for the individual. Prerequisites to nonbiased assessment are identified, including good fundamentals and ethical practices, clarification of purpose, and multifactored assessment. Research on the System of Multicultural Pluralistic Assessment are reviewed, followed by a discussion of assessment of adaptive behavior with mentally retarded persons. Requirements of P.L. 94-142, (the Education for All Handicapped Children Act) that the child's sociocultural background and primary language be taken into account are considered. The continuing problem over defining mild mental retardation is examined. The author concludes by emphasizing the importance of multifactored assessment and of viewing nonbiased assessment as a process rather than a specific number of instruments. (CL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

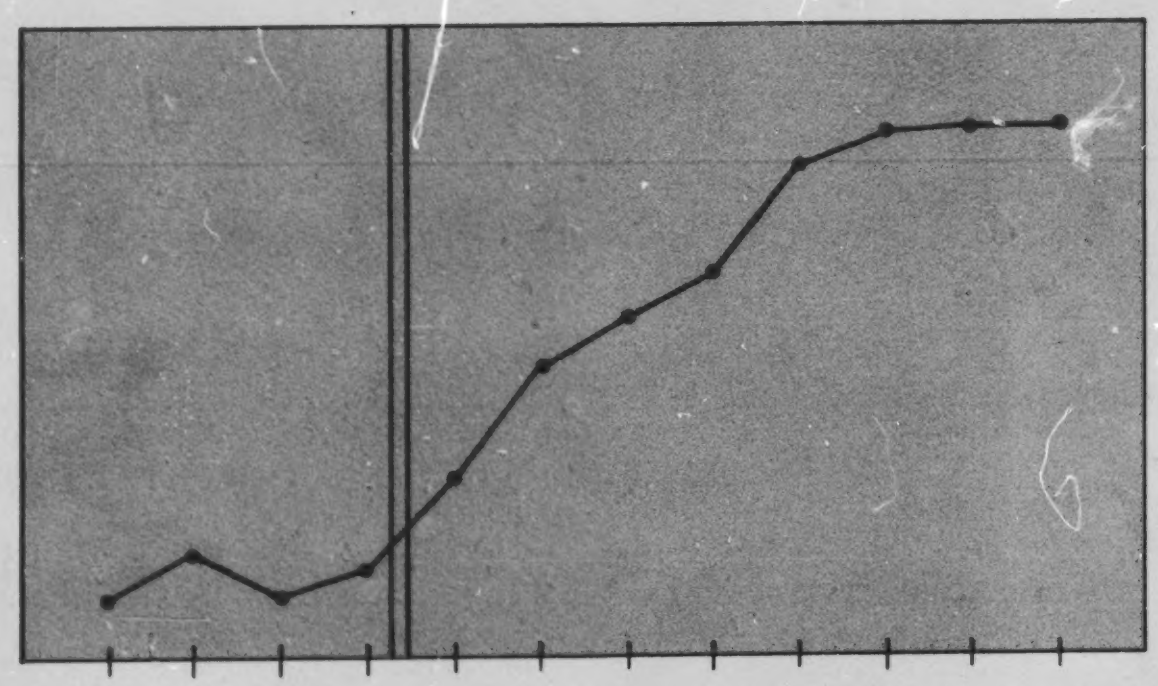
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

ED209810

NONBIASED ASSESSMENT

Daniel J. Reschly



"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Jeff Grimes
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

EC140334

December 1980

Iowa Department of Public Instruction

Iowa State University

NONBIASED ASSESSMENT

Daniel J. Reschly
Department of Psychology
Iowa State University
Ames, Iowa 50011

Portions of this paper have appeared, or will appear, in the following chapters: Reschly, 1979, Reschly, 1980, and Reschly, 1981. Permission to reproduce content from this paper is restricted to non-profit uses such as in-service training in educational agencies.

December, 1980

State of Iowa
DEPARTMENT OF PUBLIC INSTRUCTION
Grimes State Office Building
Des Moines, Iowa 50319

STATE BOARD OF PUBLIC INSTRUCTION

Susan M. Wilson, President, Waterloo
Karen K. Goodenow, Vice President, Wall Lake
William N. Cropp, Des Moines
Jolly Ann Davidson, Clarinda
Stephen C. Gerard, Sigourney
Dianne L.D. Paca, Garner
Mary E. Robinson, Cedar Rapids
John E. van der Linden, Sibley
Harold R. Yeoman, Monticello

ADMINISTRATION

Robert D. Benton, State Superintendent and Executive Officer
of the State Board of Public Instruction
David H. Bechtel, Administrative Assistant
James E. Mitchell, Deputy State Superintendent

DIVISION OF SPECIAL EDUCATION

J. Frank Vance, Director
Richard E. Fischer, Chief, Support Services
Jeff Grimes, Consultant, School Psychology

NONBIASED ASSESSMENT

Table of Contents

THE PROBLEM 1

LEGAL REQUIREMENTS 2
- Litigation 2
- Implicit Issues in Litigation 3
- Legislation 7

CONCEPTS OF BIAS IN ASSESSMENT AND RESEARCH WITH THE WISC-R 7
- The Many Definitions of Bias 7
- Construct Validity/Content Bias 9
  - Mean Differences 10
  - Item Bias 12
  - Psychometric Characteristics 14
  - Factor Analysis 15
  - Summary 16
- Atmosphere Bias 16
- Bias in Test Use: Predictive Validity 17
- Bias in Test Use: Social Consequences 19
  - Overinterpretation, Misinterpretation and Misuse 20
  - Selection Ratios 23
- Summary 24

OUTCOMES CRITERION 24

PREREQUISITES TO NONBIASED ASSESSMENT 26
- Good Fundamentals and Ethical Practices 26
- Clarification of Purpose 26
- Relevant Assessment 27
- Placement Options and Effective Programs 29
- Multifactored Assessment 29

THE SYSTEM OF MULTICULTURAL PLURALISTIC ASSESSMENT 32
- Mercer's Riverside Studies 32
- SOMPA Models and Measures 34

ADAPTIVE BEHAVIOR 35
- Adaptive Behavior and Definitions of Mental Retardation 35
- Conceptions of Adaptive Behavior 36
- Assessment of Adaptive Behavior 37
- Adaptive Behavior Measures for the Mildly Retarded 39
- Research on Adaptive Behavior 43
- Unresolved Issues in Adaptive Behavior 45
- Combining Intelligence and Adaptive Behavior Data 46

SOCIOCULTURAL BACKGROUND 50
- The Concept of Eth-Class 50
  - SOMPA Sociocultural Scales 51
  - SOMPA Estimated Learning Potential 52
  - Meaning of ELP 52

PRIMARY LANGUAGE 54

MILD MENTAL RETARDATION: "A CONTINUING DILEMMA" 55
Definitions and Classification Criteria 55

NONBIASED ASSESSMENT: SOME TENTATIVE CONCLUSIONS 58

APPENDIX A: Protection in Evaluation Procedures Provisions 60

APPENDIX B: Northeast Regional Resource Center Outline of Nonbiased Assessment Procedures 61

APPENDIX C: Guildeines for Special Education Assessment 66

APPENDIX D: Adaptive Behavior Instruments 69

REFERENCES 70

NONBIASED ASSESSMENT

Daniel J. Reschly
Iowa State University

THE PROBLEM

When in danger, when in doubt,
Run in circles
Yell and shout.

The issues of bias in tests and in assessment have provoked high-frequency behaviors of the type suggested in the anonymous saying quoted here. Much heat has been generated through the yelling and shouting, but relatively little light. Illumination of improved practices in psychology and education, especially procedures that would expand opportunities and improve competencies for children, have been conspicuously absent in most of the discussions.

Perhaps the main difficulty stems from a focus on the wrong problems and the wrong questions in the discussion of nonbiased assessment. The major concern has been with the assessment of minorities, particularly questions related to whether specific tests are biased or unfair when used with black, Latino, or Native American children. The issues related to the use of tests with children from minority backgrounds are legitimate and important to raise. However, a more significant issue to address is whether we can ensure educational experiences that maximize competencies and opportunities for minority students.

Several of the wide assortment of definitions and criteria for determining bias in tests or assessment are discussed and evaluated in this paper. Although each of these conceptions has merit, a more comprehensive view of bias in assessment is proposed. Bias in tests, or bias in assessment generally, should be evaluated according to the criterion of outcomes for individuals. The concern for outcomes for individuals directs our efforts toward ensuring that assessment activities yield information useful for educational and psychological interventions, and toward the effectiveness of these interventions.

Effective solutions to the challenges posed by nonbiased assessment will not be found simply in new tests or revisions in present tests. There are no culture free or fair tests! Better assessment will be part of an effective response, but this alone is not the answer. Further, other solutions such as scrupulous avoidance of overrepresentation of minorities in special education may satisfy certain external agencies, but this too is an ineffective solution.

Effective solutions are possible only through recognition of the larger problem. The critical issue is the quality and effectiveness of the educational services provided to economically disadvantaged students. Our part of the problem as special education and related services personnel is the quality and usefulness of special education services provided to economically disadvantaged persons referred to special services.

It is this group, i.e., economically disadvantaged students referred for special services, that has received an enormous amount of attention in recent years. The discussions have been heated and controversial. Economically disadvantaged students, often with minority status, have been and are placed in special education programs at a rate that is disproportionate to their numbers in the total population. This overrepresentation has been the subject of extensive litigation, legislation, and Federal Office for Civil Rights activities.

Understanding the litigation, legislation, and federal compliance activities is

important in developing effective responses to the challenge of nonbiased assessment. The implicit assumptions in the litigation must be understood in order to establish open communication, and to identify practices in need of reform. Finally, clarification of the implicit assumptions leads to the view of nonbiased assessment as a process rather than a magic test or simple avoidance of overrepresentation.

LEGAL REQUIREMENTS

Bersoff (1979) has provided a comprehensive review of the evolution of judicial examination of psychological assessment. The courts have had an enormous influence on psychological assessment and special education services. Nearly all of the major principles codified in legislation of the mid and late 1970s appeared earlier in judicial opinions or consent decrees (Turnbull, 1978). Appropriate assessment and appropriate educational services for economically disadvantaged minority students were among the most important issues in litigation in the early 1970s.

Litigation

Diana and Guadalupe Cases. Two cases in the early 1970s involved nearly identical issues concerning psychological assessment and special education services for the mildly retarded. The Diana (Diana v State of California, 1970) and Guadalupe (Guadalupe v Tempe Elementary District, 1972) cases were filed as class action suits on behalf of minority/bilingual students placed in programs for the Educable (mild) Mentally Retarded (EMR). In both cases plaintiffs presented evidence indicating overrepresentation of minority/bilingual students in EMR programs. For example, in the Diana Case the enrollment of Hispanic students in Monterey County California was 18.5% of the total enrollment, but one-third of the students in EMR special classes were Hispanic. This overrepresentation was viewed as promoting segregation and in violation of Fourteenth Amendment rights to equal protection of the laws. Conventional psychological assessment practices, particularly intelligence tests, were regarded by the plaintiffs (and apparently the courts) as the major cause of the overrepresentation.

Both cases were resolved through consent decrees negotiated between plaintiffs and defendants, and then approved by the courts. The consent decrees specified a number of reforms in psychological assessment practices including the following: Assessment of primary language competence, and administration and interpretation of tests in a manner consistent with the child's primary language; emphasis on nonverbal or performance tests in classification decisions with bilingual students; and immediate reevaluation of students who may have been misplaced. In addition the Guadalupe consent decree lowered the IQ cut off for classification/ placement decisions; required assessment of adaptive behavior outside of school; and required that intelligence test results not be the exclusive or primary basis for classifying children as mildly retarded in the public schools. Implicit in both cases were the assumptions that intelligence tests, especially verbal tests, were biased against bilingual students and that special class programs for the mildly retarded were ineffective and stigmatizing.

Larry P. v Riles (1972, 1974, & 1979). The Larry P. case was a class action suit related to the basic issue of overrepresentation of minority students in programs for the mildly retarded. Larry P. was filed on behalf of Black children placed in programs for the mildly retarded. The case was filed originally in November, 1971; an injunction was issued by the Federal District Court for Northern California in June, 1972; an expanded injunction was issued in 1974; the case was in trial from October, 1977 to May, 1978; and an opinion was issued by Judge Peckam in October, 1979. The Larry P. case has already been before the courts for nearly a decade. Appeal of the decision perhaps to the U.S. Supreme Court is considered likely. The Larry P. trial generated a 10,000 page transcript much of which came from expert witnesses for the plaintiffs and defense.

The preliminary injunction in Larry P. restrained the defendents (officials of the San Francisco Public Schools and the California State Department of Education) from

> "placing Black students in classes for the educable mentally retarded on the basis of criteria which place primary reliance on the results of IQ tests as they are currently administered, if (emphasis added) the consequence of use of such criteria is racial imbalance in the composition of such classes" (Larry P. v Riles Court Injunction, 1972).

In 1974 the plaintiffs obtained an expansion of the injunction to all school districts in California. The 1979 court opinion also placed a ban on the use of intelligence tests with black students. The key statement in the decision which was 131 pages in length was

> "Defendants are enjoined from utilizing, permitting the use of, or approving the use of any standardized intelligence tests, ..., for the identification of black E.M.R. children or their placement into E.M.R. classes, without securing prior approval by this court" (Larry P. v Riles, 1979, p. 104).

The implications of the Larry P. opinion for school psychology and special education are unclear, but potentially enormous (See School Psychology Review, Vol 9, No. 2, 1980). In my view, the court, through the plaintiffs actions, identified a significant problem; namely, the appropriateness of segregated special classes for "six hour" retarded children. The opinion, however, is an instance of Right Problem-Wrong Solution. A number of underlying assumptions are apparent in Judge Peckam's opinion. These assumptions and the issues they represent are probably more important in developing solutions to the problem of appropriate education for all children than the narrow issue of potential bias in IQ tests (see later sections).

PASE vs Hannon (1980). A recent decision from a Federal District Court in Illinois addressed the same issues as previous placement bias cases, but reached a markedly different decision. Again, the primary issue was alleged bias in intelligence tests. In contrast to previous decisions such as Larry P., the judge concluded that very few items on conventional tests were biased and that other sources of information were just as important as test scores in classification/placement decisions.

In view of the recent PASE Opinion, and the expected appeals in both PASE and Larry P., the present legal situation is highly ambiguous. Appeals typically are very time consuming. Both cases may reach the U.S. Supreme Court in the mid to late 1980s. Resolution of the question of bias through the courts has not been possible to date for many reasons. Perhaps the most important reason is the inconsistent conceptions and evidence on bias as well as the inherent nature of the judicial process (See later sections).

Implicit Issues in Litigation

Although the litigation concerning overrepresentation of minorities in special class programs for the mildly retarded focused on alleged bias in intelligence tests, a number of implicit assumptions were made by the plaintiffs and accepted by the courts. These assumptions represent unresolved issues in the professional literature, and are more important to the provision of fair and effective services to children than the

narrow (and perhaps unresolvable) issue of bias in intelligence tests. Examination of these assumptions provides a better perspective on recent legislation as well as suggestions for different approaches to the problems of bias in assessment and appropriate classification/placement decisions with minority students.

Nature-Nurture. The debate over the relative effects of heredity and environment in determining intelligence predates the development of measures of intelligence. This very old debate has not been resolved and is not likely to be resolved in the forseeable future. The controversy was increased dramatically in the 1970s with the extension of the hereditarian view to explain differences between racial groups (Jensen, 1969). The views of other participants in the debate were sometimes inflammatory (Shockley, 1971), and interpreted as stemming from frankly racist motives. The reaction of many, including psychologists and the lay public, black as well as white, was to denounce these widely publicized theories and interpretations of equivocal evidence. Of particular importance to school psychologists were the vehement attacks on intelligence tests that were prompted by the suggestions that racial differences were due to hereditary factors. These reactions appeared in the literature, and were undoubtedly a principal factor leading to the Larry P. and other court actions. The available evidence on differences among races is equivocal. Strong hereditarian, strong environmental, or interactionist positions have been supported by citing evidence (Loehlin, Lindzey, & Spuhler, 1975). The debate is therefore unlikely to be resolved through conventional empirical methods. The alternative apparently chosen by the critics was to force a kind of resolution through legal procedures. The ban on the use of ability tests with black students in California from the Larry P. decision might be extended to other locations and to other minority groups. Use of intelligence tests with minorities might be severely restricted in the future, but it is unlikely that even this radical step would end or lead to resolution of the nature-nurture debate. Moreover, eliminating the use of ability tests with minorities would accomplish little if anything toward elimination of existing barriers to the full participation of all persons in the economic and social order, and would likely be counterproductive in that effort. Although not mentioned explicitly, the nature-nurture issue was a crucial factor in the Larry P. litigation.

Meaning of IQ Test Results. A number of myths regarding the meaning of intelligence test results have been around for several decades. Of particular concern are the beliefs that IQ test results are predetermined by genetic factors, that intelligence is unitary and is measured directly by IQ tests, and that IQ test results are fixed. The available evidence clearly refutes these myths (Hunt, 1961; Reschly, in 1980), and the vast majority of professional psychologists do not harbor misconceptions. Kaufman (1979a) provided an excellent discussion of the assumptions underlying and the meaning of intellectual assessment. His views are probably typical of most professional psychologists. However, many consumers of IQ test results such as teachers, parents, and the lay public generally hold these misconceptions. Recent suggestions to change the term IQ to School Functioning Level (Mercer, 1979) or Academic Aptitude (Reschly, 1979) are designed to reduce these misconceptions.

A significant portion of the testimony in Larry P. was devoted to disproving these myths. This testimony has a "straw man" quality. The fact these myths were an implicit issue in the litigation provides further evidence for the need to clarify the meaning of IQ test results, and perhaps, renaming the construct.

Labeling Effects. Implicit in all of the litigation was the assumption that classification as Educable Mentally Retarded was stigmatizing and humiliating with probable permanent effects. The controversy over labeling is far from resolved. The available empirical evidence does not support the self-fulfilling prophecy notion and

direct effects of labels on the behavior of children or adults have been difficult to document (MacMillan, Jones, & Aloia, 1974). The dilemmas associated with classification have been prominent in the special education and school psychology literature for the past decade. Much of the discussion has not been guided by empirical data. The dilemma was described well by Gallagher (1972) who acknowledged the inevitability of classification, but suggested that the crucial factor was whether the benefits of services provided as a result of the label were sufficient to justify the possible risks of the label. This risks/benefits criterion should guide our efforts in the future to deal with this issue.

Meaning of Mild Mental Retardation. The reasoning of the Larry P. decision was that the plaintiffs were not "truly retarded" despite low IQs, low academic achievement, and teacher referral. The effort to identify "true" mental retardation appears to be related to confusion of mild with the more severe levels of mental retardation. The criteria for "true" mental retardation are apparently believed to require comprehensive incompetence, permanence, and evidence of biological anomaly (Mercer, 1973; 1979). In contrast, the AAMD classification system does not specify etiology or prognosis. In addition, different domains of adaptive behavior are emphasized depending on the age of the individual. There was little doubt that the plaintiffs in the placement litigation had serious academic problems. The question was whether they were "truly" retarded, or whether they merely performed within the retarded range due to biases in the IQ tests. Confusion over the meaning of mild mental retardation and questions concerning the criteria for adaptive behavior were key issues in the cases (See later section).

Efficacy of Special Classes. The efficacy of special classes for the mildly retarded was challenged forcefully in the 1960s (e.g., Dunn, 1968). The lack of clear evidence to support the effectiveness of special classes along with the allegations concerning the negative effects of labels created a difficult situation for the defendants (school districts and state departments of education) in the placement litigation. Further, the overrepresentation of minorities in segregated special classes raised questions about segregation of student groups by race. In several instances the school districts and state departments of education did not defend their programs in court. Consent agreements were negotiated out of court. In the Larry P. case a defense of the programs was attempted, but unsuccessfully. It should be noted that if the special class educational programs were as poor as alleged, then no child regardless of race or social class should be placed in such programs. The crucial issue, but implicit in the litigation, was effectiveness of special class programs. Unfortunately, the plaintiffs and courts seemed to focus on the criteria for placement of students rather than the effectiveness of the programs as such. Additional research on the effectiveness of special education programs using longitudinal designs is clearly needed.

Meaning of Bias. Many definitions of bias in tests have been proposed in the psychological and educational measurement literature (see later section, this paper). Two criteria for bias have been implicitly accepted by the courts.

In all of the placement litigation the plaintiffs presented evidence on overrepresentation of minorities in special education programs. The overrepresentation data bear closer analysis. In many cases these data may have been misunderstood. For example when the original Larry P. case court injunction was expanded in 1974, the percentage of black students in the San Francisco schools was 30%, but the enrollment in programs for the Educable Mentally Retarded (EMR) was 60% black. The comparable state wide figures in California for the past ten years have been approximately 10 and 25 per cent respectively for total black student enrollment and black student enrollment in EMR special classes. Quite clearly, black students have been overrepresented. However, these figures have sometimes been understood to indicate

that many if not a majority of blacks were diagnosed as mentally retarded allegedly because of biased IQ tests. For example, in the Larry P. opinion these percentages were characterized as being "grossly disproportionate" (p. 94) and as indicating "overwhelming disproportions" (p. 101). However, over the past decade the percentage of the total black student population in California placed in EMR special classes has varied from about 3.2% in 1968-69 to about 1.1% in 1976-77 (Reschly, 1980). Only a very small percentage of minorities have been placed in special class programs even in cases that appear to reveal very high overrepresentation. These data certainly do not support the notion that IQ tests have a pervasive deleterious effect on black children.

The possible causes of the overrepresentation and other factors associated with overrepresentation should also be considered. The overrepresentation of males in programs for the mildly handicapped (i.e., learning disability, behavior disorders, and mild mental retardation) is greater than the overrepresentation of minorities. The overrepresentation of students from economically disadvantaged homes is even more pronounced for the category of mild mental retardation regardless of racial/ethnic status. Minority status and socioeconomic status are (unfortunately) not independent. The intriguing question is whether minorities are overrepresented beyond the level that might be predicted from socioeconomic status data.

Overrepresentation is a simplistic and often misunderstood notion of test bias. Nevertheless this definition continues to be used by the courts (e.g., Mattie T. vs Holladay, 1979). If carried to its logical conclusion this definition could result in elimination of virtually all special services programs due to alleged race/ethnic, sex, or socioeconomic bias. This illogical outcome is not in the best interests of children.

The other definition of bias used by the courts is mean differences in scores among groups. This definition is discussed in a later section of this paper.

Special education placement litigation has been a significant influence in recent years. Unfortunately, the courts by their nature are not a desirable mechanism for resolving disputes in the behavioral sciences. In contrast to the behavioral sciences and professions, the fundamental purposes and methods of resolving issues are quite different in the courts. The legal system in the placement litigation is concerned with abstract principles of justice, particularly as they apply to groups of persons. The sciences are devoted to "truth" which is recognized as being tentative and approximate. The perspectives of professional personnel such as special educators and school psychologists are typically focused on the individual who is having significant learning or behavioral problems in the classroom. The explicit and implicit issues in the litigation are at best ambiguous. None of the issues can be resolved unequivocally through the scientific method of theory, research, and analysis of data. The available evidence is at the level of probability statements which would justify decisions using language such as "might" or "should." The professionals involved appear to operate in a manner consistent with this evidence. For example not all children with low IQ scores are placed in special programs, and a few with IQ scores above cut off criteria are placed. These decisions are based on a comprehensive view of the individual and the best estimates of what is best for that individual. The overrepresentation that has resulted has been the culmination of decisions about individuals, not decisions about groups. The status of groups of persons has been an important area of judicial inquiry which has expanded since the Brown decision in 1954. However, the courts by their nature reach decisions which pertain to groups and are stated in decisive, unequivocal language such as "shall" or "must." The court remedies are therefore rarely consistent with the scientific evidence, or the approach of professionals.

Legislation

The litigation of the late 1960s and early 1970s was an important source of influence on State and Federal legislation in the mid 1970s. The appendices to this paper include the Protection in Evaluation Procedures Provisions (Sections 121a. 530-121a. 534, Federal Register, 1977) of PL 94-142. These requirements are particularly relevant to the challenge of providing appropriate assessment services for all students, and should be read carefully by all special education personnel. Two provisions are particularly important.

> "Testing and evaluation materials and procedures used for the purposes of evaluation and placement of handicapped children must be selected and administered so as not to be racially or culturally discriminatory." Section 121a. 530(b), Federal Register, 1977.
>
> "In interpreting evaluation data and in making placement decisions, each public agency shall: Draw upon information from a variety of sources, including aptitude and achievement tests, teacher recommendations, physical condition, social or cultural background, and adaptive behavior; Insure that information obtained from all of these sources is documented and carefully considered." Section 121a. 533(a, 1 & 2), Federal Register, 1977.

The requirement that assessment be nondiscriminatory is deceptively simple. The language is unequivocal, but no definition is provided and no criteria are available in the legislation or rules and regulations concerning implementation of the regulation. The apparent solution was to require that a broad variety of information be considered including social or cultural background and adaptive behavior. The meaning, measurement, and use of these concepts are also far from clear.

CONCEPTS OF BIAS IN TESTS AND RESEARCH WITH THE WISC-R

Much of the special education placement litigation as well as other discussions of overrepresentation have assumed that conventional tests are biased against minority students. Careful examination of the educational and psychological literature reveals a different picture. There are many definitions of bias, a variety of ways to analyze the data, and widely varying conclusions reflected in this literature. Surprisingly, some of the widely held assumptions about common tests simply are not supported by empirical evidence.

The Many Definitions of Bias

The concept of test bias has been defined in many different ways in the recent literature. In what is perhaps the most comprehensive discussion of different definitions Flaugher (1978) identified eight separate concepts of bias in tests. Other recent examinations of test bias have analyzed the different values which underlie varying positions (Hunter and Schmidt, 1976); the different procedures for enhancing fairness vs. social equity in selection (Petersen and Novick, 1976); and the different outcomes of empirical examination of test bias depending on the definition and criteria used (Reschly, 1981). Others, e.g., Ysseldyke (1979), have stressed factors such as naturally occuring pupil characteristics (e.g., physical attractiveness) which bias decisions before and after formal assessment activities are conducted.

Close examination of some of the recent definitions of bias will reveal both the

varying conceptions and criteria as well as some common features. Flaugher (1978) identified the following definitions of test bias as mean differences, overinterpretation, sexism, differential validity, content, selection model, wrong criterion, and atmosphere (referring to situational factors in examiner characteristics, examiner-examinee interaction, etc.). Jones (1978) suggested that test bias might exist at the content level in the selection of items; in the standardization where decisions are made concerning the population for whom the test is appropriate; in the administration of the test where the examiner may be unfamiliar with the culture of the child; and in the validation where research may not be conducted concerning test validity for culturally different persons. Mercer (1978) suggested that the following five lines of evidence establish the existence of bias in tests: Test items from a single cultural heritage; Differences in average scores among different racial and cultural groups; Sociocultural differences within and between cultural and racial groups with these differences accounting for a significant proportion of the variance in test performance; Experimental studies demonstrating the effects on test performance of early interventions with culturally different children; and The effects of adoption of minority children into core culture homes.

It is interesting to note the commonalities in the definitions proposed by these diverse authors who represent different disciplines, cultural groups, and perspectives. The criterion of item or content bias is mentioned by all with at least two of the authors agreeing on the criteria of average score differences (Flaugher and Mercer), administration or atmosphere effects (Flaugher and Jones), and differential validity (Flaugher and Jones). Similar concerns appear to be expressed, although in different ways, by Flaugher and Jones regarding misinterpretation and the appropriateness of the criterion in validity studies.

An obvious question is which definition of test bias is correct? What criterion should we use in the examination commonly used tests? The answer, perhaps unfortunately, is far from simple. Flaugher (1978) argues that all of the definitions are "right" in the sense that test bias is a public concern, i.e., not restricted to an academic discipline, and significant numbers of citizens have legitimate interests and concerns in the definition used.

Table 1 provides a list of the common definitions of test bias, and a summary of results from many studies. On most criteria, conventional intelligence tests are not found to be biased. However, the social consequences of test use have often been negative, an issue to which we shall return in this paper.

Table 1

Summary of Concepts and Empirical Studies of Bias in Tests

Definition of Bias	Empirical Studies	Results Confirmed/Equivocal/Not Supported
1. MEAN DIFFERENCES	Large number of studies.	Economically disadvantaged minority students obtain lower average scores. The size of the differences vary by group and/or for some groups, by type of measure.
2. ITEM BIAS	Several recent studies using the WISC-R. Many studies with group tests.	Subjective judgments usually identify many items as biased. However, subjective judgments are unreliable. Empirical studies generally do not support the existence of item bias on conventional tests.

Definition of Bias	Empirical Studies	Results Confirmed/Equivocal/Not Supported
3. PSYCHOMETRIC	Several recent studies.	Psychometric characteristics such as reliability, item x total, subtest x scale, etc. are the same regardless of group.
4. FACTOR ANALYSIS	Several recent studies.	The factor structure on tests such as the WISC-R is largely the same regardless of group.
5. ATMOSPHERE BIAS	Many studies.	Inconsistent results, often contradictory. The size of the effects; if real, is small.
6. PREDICTIVE VALIDITY TESTS OF ACHIEVEMENT	Many studies.	The relationship between ability and achievement tests is virtually the same regardless of group. Issue of "autocorrelation" is unresolved.
7. PREDICTIVE VALIDITY TEACHER RATINGS/ GRADES	Few studies.	Inconsistent results, apparently due to type of criterion measure.
8. SOCIAL CONSEQUENCES Misuse, misinterpretation, overinterpretation	Few published studies, considerable anecdotal and historical evidence.	Conventional tests are frequently overinterpreted and/or misinterpreted. Test results have been used to justify restrictive and sometimes racist social policies.
9. SELECTION RATIOS	Many "indirect" studies.	Economically disadvantaged, minority students are overrepresented in special education programs for the mildly retarded. Tests are used as part of that process. Whether test use increases OR decreases the overrepresentation is unclear.

CONSTRUCT VALIDITY/CONTENT BIAS

Perhaps the most commonly used definition of test bias is the assertion that the WISC-R and other conventional standardized tests measure a different attribute when used with non-Anglo persons. This assertion amounts to a criticism that the construct validity of the test is not the same for all groups. If the test measures different attributes and the items function differently depending on group membership, then the meaning and usefulness of the test results probably are diminished. Moreover, the mean differences between groups are then attributed to inappropriateness of the test items, and other explanations such as economic disadvantage are rejected. Thus, the construct validity/content bias conception has broad implications for examination of bias in specific instruments such as the WISC-R. A number of different criteria have been suggested for examination of construct validity/content bias. Data from studies

using the WISC-R will be discussed in relation to four criteria: mean differences, item bias, psychometric characteristics, and factor analysis.

Mean Differences.

That different sociocultural groups obtain higher or lower scores on the average on standardized ability and achievement tests is one of the most controversial and oldest observations in the history of psychological measurement. A variety of attempts have been made to eliminate these differences through changing the nature of the tests. All such efforts to produce culture-free or culture fair tests have failed if for no other reason than the fact that the concept of culture free or fair is seriously flawed. Anastasi (1976) points out that the entire notion of assessment is culture bound from the beginning steps of specifying what to measure to the final steps of gathering validity data on the relationship of test results to some criterion behavior. The very practical question is, how would we use a culture free test even if one could be developed?

Although eliminating mean differences through revision of the tests seems virtually impossible, knowledge of the nature of the mean differences is important information for test use and interpretation of test results. Mean differences and variations in patterns of performance among groups are not simply due to the factor of socioeconomic status (SES). For example, Lesser, Fifer, and Clark (1965) reported that most of the differences in level of performance among groups could be explained by SES. However, sociocultural group had a significant influence on the pattern of performance independent of SES. As we shall see this finding appears to hold true for the performance of different groups on the WISC-R.

In Table 2 the means for different groups on the WISC-R are reported from data obtained from four studies of random samples.

From the somewhat limited and controversial perspective of test bias as mean differences among groups, the WISC-R would be regarded as biased against Black, and Hispanic groups. In all studies which included Blacks and Hispanics, both groups obtained significantly lower scores on the WISC-R. Some variation in pattern of performance is also apparent in these data. Hispanic students obtained significantly higher scores on the Performance Scale than on the Verbal Scale. It might be noted, however, that the Performance Scale scores generally were still below the population average, and that not all non-Anglo groups score higher on the Performance Scale. Cautious use of these results in generalizations to groups from other regions clearly is necessary.

Data on the existence of mean differences do not of course provide any information on causation. However, average scores below the population mean are not restricted to non-Anglo groups. Identifiable groups of White Anglo Saxon Protestants also score below population means, e.g., Appalachian Whites. This fact should certainly convince any remaining skeptics that it is not race or ethnicity per se that accounts for all of the differences among groups. In the recent literature, three plausible explanations have been proposed to account for the differences among groups. Trotman (1977) provides a good overview of the explanations of cultural differences, cultural disadvantage, and genetic inferiority. Analysis of the logic and data for each of these explanations has consumed enormous amounts of time and space in the psychological literature. Each of the explanations has obvious implications for interpretation of the WISC-R, and for social policy. For example, the cultural differences explanation of the variations in group means would suggest that the content of the WISC-R does not reflect what is regarded as intelligent behavior in non-Anglo (non-middle class) cultures. The data which follows on item bias are at least partly relevant to this point

Table 2

Mean WISC-R Scores for Different Sociocultural Groups

Group	Anglo			Black			Hispanic		
Sample Scale[1]	V	P	FS	V	P	FS	V	P	FS
1. Kaufman & Doppelt (1976) Standardization Sample	102 (N=1870)	102	102	88 (N=305)	87	86	Not available		
2. Mercer (1979) SOMPA Standardization Sample (CA)	102 (N=604)	104	103	89 (N=456)	90	88	88 (N=520)	98	92
3. Reschly (1978) Pima Co. (AZ)	101 (N=252)	102	101	86 (N=235)	89	86	85 (N=223)	93	88
4. Reschly & Ross-Reynolds, 1980 Iowa Assessment Project	108 (N=100)	110	110	96 (N=100)	96	95			

[1] V = WISC-R Verbal Scale IQ Score
P = WISC-R Performance Scale IQ Score
FS = WISC-R Full Scale IQ Score

of view. The social policy implications of the cultural difference view might be to eliminate conventional tests, to develop culturally specific tests (Williams, 1971), or to correct the bias in current tests (Mercer, 1979). The cultural disadvantage view would emphasize the inadequate stimulation for intellectual development in lower social class homes. The solution to differences in group means from this perspective would be to provide intensive early interventions (Garber, 1975) and compensatory education programs. The current tests such as the WISC-R, however, are accepted as valid indicators of intellectual ability (scholastic aptitude). In the past decade the debate over the third explanation, genetic inferiority (Jensen, 1969), has generated enormous controversy. The data available currently, and the kind of data that can be generated, provides an inadequate basis for resolution of the question of hereditary differences among groups. A complete review of the data on the nature-nurture issue is far beyond the scope of this paper. The interested reader is encouraged to examine Brody and Brody (1976), Jensen (1973), Loehlin, Lindzey, and Spuhler (1975), and Samuda (1975). Perhaps the most objectionable feature of some versions of the hereditarian position is the recommendation of changes in social policy as a result of data which are at best tentative. The sense of outrage among minority psychologists and the efforts to ban tests can perhaps be understood if we are aware of some of the extreme hereditarian views, e.g., Shockley (1971). Other implications of the hereditarian view are to place less emphasis on governmental support of early intervention and compensatory education programs, and perhaps unintentionally, but implicitly, more emphasis on interpretation of IQ scores as reflecting the genetic endowment of the individual. The proper interpretation of IQ test scores will be addressed in a later section.

Item Bias.

Allegations of cultural bias in the items used on conventional tests have been and continue to be the most popular of the criticisms of standardized tests. In fact examination of an item from a current standardized test to support the allegation of bias in all of the items appears to be an increasingly popular indoor sport. Examples of subjective judgments of item bias are numerous (e.g., APA Monitor, 1977; Dent, 1976; Williams, 1971). The implicit assumption is that all items on the test are biased if one or a few of the items are apparently biased. If the test is presumed to be biased on the basis of inappropriate items, then the test results are presumed to be "inaccurate" and unfair. If the items are biased, usually meaning that opportunity to learn the content of the item is not common to all environments, then the test results certainly do not reflect, and cannot be interpreted as evidence of "innate" intelligence. However the IQ test results are not direct measures of innate ability for any group.

The distinction between cultural bias and cultural loading is important to this discussion. The degree of cultural loading of an item, that is, the likelihood of success on the item for persons with different backgrounds and experiences, varies on a continuum. At one end of the continuum are items that could only be answered correctly by persons with highly specific backgrounds and experiences. An example might be an item that asks "Name three presidents of Iowa State University over the past century" (the present author can name only two). The item is similar to those on many intelligence tests in terms of the type of thinking required. However, only a very limited sample of persons would have an opportunity to be exposed to this information and thereby answer the item correctly. The item reflects a very high degree of cultural loading and would be regarded by most as culturally biased (as well as trivial). Some items on current standardized tests require similar kinds of thought patterns and also vary in degree of cultural loading. Another item parallel to the above example would be "Name two presidents of the United States since 1960." This item is certainly lower in cultural loading because the information required is more general, and many more persons would have an opportunity to learn the correct responses. however, some persons might still judge this item as culturally biased since the opportunity to learn the information might vary among different groups. The question of bias in an item should be determined by empirical analysis of responses from persons representing different groups, not by judgment alone.

The degree of cultural loading of an item depends on the generality of the information and the characteristics of the persons taking the test. These points are illustrated well in the development of "counterbalanced" or culturally specific intelligence tests (e.g., Dove, undated; Williams, 1975). These tests require highly specific information that is usually possessed only by persons with particular backgrounds or experiences. In Table 3 examples of culturally specific items are provided.

The evidence on item bias has been produced through two markedly different methods of examining test items. The most common method has been subjective judgments of item content. A less common method has been empirical analysis of the item responses of examinees from different racial or ethnic groups. The results of "tests" of item bias vary dramatically depending on which method is used.

Subjective judgment as a method usually involves obtaining opinions from expert representatives of the minority culture regarding whether or not the items in a test are biased against examinees from that culture. Two Verbal Scale items on the WISC-R have been cited frequently as biased. The Information subtest item, "Who discovered America?" and the Comprehension subtest item "What is the thing to do if a boy (girl)

Table 3

Sample Items From Culture-Specific Tests

Dove Counterbalanced Intelligence Test (Source Unknown)
(Urban Black Culture)

1. "T Bone Walker" got famous for playing what? (a) Trombone (b) Piano (c) T-Flute (d) Guitar (3) "Hambone".

2. A "Gas Head" is a person who has a _____. (a) Fast moving car (b) Stable of "lace" (c) "Process" (d) Habit of stealing cars (e) Long jail record for arson.

3. If you throw the dice and "7" is showing on the top, what is facing down? (a) "Seven" (b) Snake eyes (c) "Boxcars" (d) Little Joes (e) "Eleven".

4. Cheap "Chitlins" (not the kind you purchase at a frozen-food counter) will taste rubbery unless they are cooked long enough. How soon can you quit cooking them to eat and enjoy them? (a) 15 minutes (b) 12 hours (c) 24 hours (d) 1 week (on a low flame) (e) 1 hour.

5. "Jet" is _____. (a) An "East Oakland" motorcycle club (b) One of the gangs in West Side Story (c) A news and gossip magazine (d) A way of life for the very rich.

Counterbalanced Intelligence Test (Source Unknown)
(Urban Hispanic, Southwest)

1. The name "Jesus" in particular seems to disturb teachers and is nearly always changed to _____.

2. The Spanish Language spoken in the southwestern states is known by Mexicans as _____.

3. Who was considered the Mexican Robin Hood of California? _____.

4. Complete the following rhyme?

   Pancho Villa
   mato su tia
   con una tortilla
   _____

5. The first Chicano to have a big hit record was the person who sang Donna what was his name? _____.

Fry Bread IQ Test (Deer, 1980)
(American Indian Intelligence Test)

1. Social dancing and singing held after hours is called:
   (a) Indian two step (b) Pow-Wow (c) Forty-nine (d) Indian Rock

2. The Annual American Indian Fair and Exposition is held at:
   (a) Crow Agency, Montana (b) Gallup, New Mexico (c) Anadarko, Oklahoma (d) Pine Ridge, South Dakota

3. The largest of the American Indian tribes is the:
   (a) Navajo (b) Sioux (c) Cherokee (d) Creek

4. A food staple traditional to many Indian Tribes is:
   (a) Buffalo (b) Fry Bread (c) Commodity Cheese (d) Indian Round Steak (Bologna)

5. Who said "The only good Indian is a dead Indian?"
   (a) Col. George Custer (b) Gen. Andrew Jackson (c) Gen. John Pershing
   (d) Gen. Phil Sheridan

---

much smaller than yourself starts to fight with you?" are most often cited as biased against Chicano or Native American and Black examinees respectively. There are two major problems with the subjective judgment method of determining item bias. First, the inter-judge agreement among experts representing the minority cultures is usually quite low (Sandoval & Miille, 1980). Second, and most important, the results of empirical analysis do not confirm the subjective judgments.

Empirical analyses of item bias on a variety of tests have generally yielded equivocal or negative results regarding hypotheses of item bias (Sandoval, 1979). As Flaugher (1978) pointed out, if the phenomenon of item bias is real on conventional tests, it certainly does not account for a very large portion of the group differences. Elimination of biased items and rescoring the tests does not lead to significantly different results in the research published thus far.

The evidence, though certainly not definitive at this point, fails to support item bias as a significant explanation for the differences in mean scores among groups. Test items do vary in amount of cultural loading. Items on current tests are culturally loaded to varying degrees, as they must be if tests are to predict or evaluate important behaviors that occur only within a cultural context. Subjective judgments of item bias are not necessarily accurate, and revision of current tests either in the direction of greater or lesser cultural loading might have the undesirable effects of simultaneously increasing or maintaining group differences and reducing validity.

Psychometric Characteristics.

A large number of possible studies could be conducted on the internal psychometric characteristics of the WISC-R when used with different groups. Some of the possible analyses of interest would be comparisons across groups of internal consistency reliability, subtest intercorrelations, subtest correlations with Verbal, Performance, and Full Scale IQ scores, test-retest reliability, and intercorrelations of the Verbal, Performance, and Full Scale IQ scores. To date, very few such studies have been reported.

Sandoval (1979) examined the internal consistency reliability (Cronbach Alpha) of the WISC-R subtests and IQ scales using the SOMPA standardization data. The reliability of the subtests and IQ scales was high and nearly the same for Anglos, Blacks and Chicanos with the exception of Object Assembly which was more reliable for Blacks (.95) than for Anglos and Chicanos (.79 and .75, respectively). All other differences in reliability on WISC-R subtests were negligible with no systematic pattern of group differences. The reliabilities of the Verbal Scale and Performance Scale IQs were virtually identical for the three groups (rounded to .97 and .94 respectively). In the only other study pertinent to the issue of WISC-R reliability located by the author, Dean (1977) reported data on the reliability of the WISC-R from a sample of Chicano students referred for psychological evaluations in the Phoenix (AZ) area. The reliabilities of the subtests and IQ scales were comparable to the data reported in the WISC-R

manual for the age group included in the study (11.5 year olds).

Factor Analysis

Comparability of factor analytic results for different groups, and the degree to which the results of the factor analysis are consistent with the major scores and common interpretations of the test are necessary conditions for fairness in use of the test with culturally diverse persons. Indeed, if a test is not measuring the same underlying abilities or if the commonly used scores from the test represent varying abilities depending on group membership, then use of the test with culturally different persons is probably inappropriate and unfair, and the predictive validity of the test is likely to be lower for specific groups.

The appropriate number of factors that should be interpreted was a somewhat controversial issue in research on the WISC. The careful analyses of the standardization data conducted first by Kaufman (1975) and then by Silverstein (1977) seem to have resolved this problem. Both authors reported three factors; Verbal Comprehension (VC) formed by four Verbal Scale subtests, Perceptual Organization (PO) formed by five Performance Scale subtests, and a third factor labeled tentatively as Freedom from Distractibility (FD) formed by a combination of three subtests from the two scales. The reader is referred to several sources for a more thorough discussion of the use of the factor analysis results (Kaufman, 1975; 1979a; 1979b; and Reschly & Reschly, 1979).

Reschly (1978) investigated the WISC-R factor structure using data from four sociocultural groups in Pima Co., Arizona. The methodology used was a replication of Kaufman's 1975 analysis of the standardization data. The major questions addressed in this study were: The appropriate number of factors for the four groups, The comparability of the factors, The relationship of the factors to the IQ scales, and The evidence for a similar general factor among the groups.

The objective guides to the appropriate number of factors to interpret yielded inconsistent results. Three factors were indicated for Anglos, two or three factors for Chicanos depending on the criterion used, and only two factors for Blacks and Native American Papagos. In view of the contradictory evidence, both two and three factor solutions were analyzed for all four groups.

The two factor solutions were highly similar for all groups. The first and second factors for all groups conformed almost perfectly to the organization of the WISC-R into Verbal and Performance Scales. For all groups the Vocabulary (V), Information (I), Similarities (S), and Comprehension (C) subtests were the best measures of the first factor as were Object Assembly (OA) and Block Design (BD) for the second factor. Coefficients of congruence reflecting the similarity of the two factor solutions across the four groups were very high (.97 to .99).

As might have been anticipated from the preliminary data on appropriate number of factors, the three factor solutions varied significantly for the four groups. The patterns for Anglo and Chicano groups were nearly identical to the data reported for the standardization sample. The three factor solutions yielded an uninterpretable third factor for Native American Papagos, and a splitting of the major Performance Scale subtests into two factors for Blacks. In the three factor solutions the coefficients of congruence were very high scross all groups for the first factor, high for the second factor, and high and comparable only for Anglos and Chicanos on the third factor.

A final series of analyses with the Pima County data were conducted around the question of evidence for a general factor on the WISC-R for the diverse groups. Three methods of estimating the amount of variance attributable to a general factor yielded

results pointing to the conclusion that the WISC-R Full Scale IQ score reflects the same attribute regardless of group membership. The fluctuations between groups in amount of variance attributable to a general factor were minor.

Differences between groups in the factor analysis of the Pima County data were found only in relation to the nature and composition of the third factor. The meaning of this factor, which accounts for a relatively small proportion of the WISC-R variance, has never been entirely clear. The other evidence from this study clearly supports the construct validity of the WISC-R with non-Anglo groups. Nearly identical two factor solutions which conform closely to the organization of the scales were found. A large general factor was clearly apparent in about the same form and magnitude for all groups. Thus, the usual interpretation of the Full Scale IQ as an index of general intelligence (scholastic aptitude) and the Verbal-Performance scale distinction appear to be equally appropriate for Anglo and non-Anglo groups.

Summary: Construct Validity/Content Bias. In this section, data on four different methods for determining construct validity/content bias in the WISC-R were analyzed. Although the mean differences criterion is somewhat inconsistent with the other criteria discussed in this section, it was included here because mean differences are frequently explained by allegations of content bias or assertions that the WISC-R measures a different attribute in Anglo and non-Anglo groups. The mean differences criterion raises troublesome questions because it seems to lead directly to prejudging and ruling out the reality of differences among groups (Thorndike, 1971). Mean differences as such provide only weak evidence on test bias, and have no bearing on the question of bias in test use. Nevertheless, this criterion was discussed here since it is even less consistent with the other general conceptions of test bias around which much of this chapter is organized. With these limitations in mind, it is appropriate to conclude that some people would regard the WISC-R as biased since mean differences among groups do exist. However, the hypotheses that the mean differences are cuased by item bias or that the test measures a different attribute in a different fashion depending on group membership are simply not supported by data. Whatever it is that the WISC-R measures, a question to be discussed later, it appears that the same attribute is measured in the same way regardless of group membership.

ATMOSPHERE BIAS

In addition to bias in content, another frequent criticism of standardized tests is that the atmosphere of the testing situation is unfair to minority children. Two general aspects of the testing environment are mentioned most frequently as possible sources of unfairness: (1) The kinds of responses and nature of the effort required on the test or (2) The nature of the interaction with the examiner may be inconsistent with the child's background or experiences.

A great amount of research has been conducted on atmosphere bias, and is well reviewed by Sattler (1970, 1973, and 1974). The interested reader is encouraged to pursue further information in those sources. The major conclusions from this research are the following:

1) Much of the research was poorly designed.

2) Some of the studies used experimental manipulations that are atypical and inconsistent with good testing practices. For example, token reinforcers provided for correct answers.

3) The results of reasonably well-controlled studies in which the variables manipulated were within the range of good testing practices are contra-

dictory. For example, the degree of warmth, amount of encouragement, time devoted to establishing rapport prior to testing, and sex or race of examinee, have been studied with mixed results. Inconsistency is the rule rather than the exception in studies of examiner effects.

4) Examiner expectancies for performance may influence scoring of responses on items where there is some subjectivity in evaluating responses, for example, Vocabulary subtest of the Wechsler scales.

5) When differences due to atmosphere effects are reported, the size of the differences is usually fairly small.

6) If the phenomenon of atmosphere effects is real, it is doubtful that it accounts for very much of the differences among groups on standardized tests (Flaugher, 1978).

Although the research on atmosphere effects does not support the existence of this sort of bias for groups, these results do not necessarily generalize to all natural settings or to the performance of all individuals. It is essential to recognize the basic assumption of maximum effort on ability, achievement, and aptitude tests. If the child cannot or does not perform as well as possible due to unique features of the testing environment, the results of the test are inaccurate reflections of the child's thinking competencies or academic skills. In such cases, comparisons of the child's performance to that of the normative sample are inappropriate.

Professional personnel who administer tests to culturally different persons must be sensitive to individual variations in values, motivation, language, and cognitive style, all of which could influence the results of the test. One of the most important roles of the examiner in individual evaluations is to establish the kind of climate that will produce the child's maximum effort and performance. In order to be effective in this role the examiner needs to understand and appreciate the culture of the child being assessed. Additional information regarding important considerations in assessment of non-Anglo children is provided by Sattler (1974), Hynd and Garcia (1979) for Native Americans, by Bartel, Grill, and Bryen (1973) for Blacks, and by Matluck and Mace (1973) for Chicanos.

BIAS IN TEST USE: PREDICTIVE VALIDITY

A fairly common assertion is that conventional standardized tests such as the WISC-R provide low estimates of the competencies of minority group examinees. If this assertion is correct then bias or discrimination may result from use of the test in predictions of performance on various criteria (Deutsch, Fishman, Kogan, North, & Whitman, 1964). If the test is less valid for minority examinees or if the predictions from the test vary as a function of group membership, then indeed test use is less effective, and unfair or discriminatory as well if the prediction is too low. In this section evidence will be reviewed on the validity and predictive accuracy of the WISC-R when used with minorities.

Several sources of information are available on the general issue of the validity of individual and group intelligence tests. Sattler (1974) is a particularly good source of information for children's scales and Matarazzo (1972) provides an excellent review and discussion of the relationship of ability measures to a variety of criteria. Although these sources of information are adequate to answer general questions concerning validity, they may not be sufficient to meet the PL 94-142 criterion of "validated for the specific purpose for which they are used." This is especially true when different criteria are suggested or when conventional criteria are seen as inappropriate for assessing the validity of tests.

Competing points of view regarding what criterion should be used in studies of the validity of the WISC-R (and other intelligence tests) were expressed by different witnesses in the Larry P. court trial. The debate revolved around the question of whether standardized tests of academic achievement are appropriate criteria for assessing the degree to which the WISC-R predicts school performance. Abundant data do of course exist to substantiate the fairly strong, positive correlation between the WISC-R and standized tests of achievement (typical correlations from studies are in the range of .5 to .7). Mercer contended that the ability and achievement tests were measuring the same thing; that the traditional distinction between the tests was artificial; and that both ability and achievement tests shared the same kinds of biases against minorities. Mercer then suggested that grades and teacher ratings of classroom performance were the only really independent sources of information (i.e., independent of IQ) regarding performance in the academic setting, and hence, the only appropriate criteria to use in studies of the validity of the WISC-R. Others have disputed this point of view (e.g., Clarizio, 1979a; 1979b).

Regardless of which criterion is used to assess academic performance, predictive validity of the WISC-R for the criterion school achievement is clearly a necessary prerequisite to fairness in the use of the WISC-R with minorities by school psychologists. Fortunately, recent studies do appear to support the predictive validity of the WISC-R for both types of criteria of academic performance. In Table 4 results from several large sample studies are presented.

The results from several recent studies summarized in Table 4 support the validity of the WISC-R as a predictor of achievement for minority and majority groups. The magnitude of the correlations were about the same for all groups with the possible exception of Native American Papagos where the relationships were generally lower. The correlations between the WISC-R Full Scale IQ and standardized achievement test results were in the typical range of .5 to .7 for the three groups included in both studies. Goldman and Hartig (1976) published data on the relationship of the WISC to three measures of classroom performance for large samples of Anglo, Black, and Chicano students from Riverside California. The WISC was administered in 1967 with the measures of classroom performance apparently collected at varying times between 1967 and 1969. Teacher assigned grades over the next two years were collected and organized into a composite for Academic GPA. The Academic GPA was a rather strange amalgamation of grades in academic and nonacademic subjects including, "music, health, art, reading, arithmetic, math, social studies, science, language, spelling, writing, instrumental music, physical education, composition and grammar, history, geography, and foreign language" (p. 585, emphasis added). The relationship of the WISC to the "Academic GPA" measure was relatively low for all groups, but higher for Anglos (.25) than for Chicanos and Blacks (.12 and .14 respectively). Mercer (1979) reported similar results for the same groups. Again, the measure of "Academic GPA" was a rather unusual combination of grades in academic and nonacademic subjects. Other studies using teacher ratings of academic performance revealed no evidence of differential validity (Reschly & Reschly, 1979; Reschly & Ross-Reynolds, 1980; Hartlage & Steele, 1977).

Overall, studies on the relationship between measures such as the WISC-R and academic performance are generally positive. Clearly, the only evidence for lower or differential validity when the criterion for academic achievement is a standardized test comes from one sample of Native American students. For other groups, Anglos, Blacks, and Chicanos, the WISC-R predicted standardized achievement test performance equally well regardless of group membership. The data regarding the relationship of the WISC-R to teacher ratings or grades are less definitive for a variety of reasons. There is the sticky problem of the reliability (and validity) of teacher ratings. Despite the obvious problems with this criterion, there are data to support the validity of the WISC-R as a predictor of teacher ratings for different racial or ethnic

Table 4

Correlations of WISC-R Full Scale Score with Standardized Tests of Achievement and with Teacher Ratings/Grades

Sample/Group		Achievement Measure: Test Reading	Test Math	Teacher Rating or Grades
Pima Co., AZ (Reschly & Reschly, 1979)	A	.56	.55	.35
	B	.62	.51	.45
	H	.55	.50	.38
	NAP	.41	.43	.34
Austin, TX (Oakland, 1977)	A	.72	.64	NA
	B	.64	.61	NA
	H	.64	.59	NA
Riverside, CA (Goldman & Hartig, 1976)	A	NA	NA	.25
	B	NA	NA	.14
	H	NA	NA	.12
Riverside, CA (Mercer, 1979)	A	NA	NA	.44 (Mdn.)
	B	NA	NA	.27 (Mdn.)
	H	NA	NA	.24 (Mdn.)
Iowa Assessment Project (Reschly & Ross-Reynolds, In Press)	A	Not Analyzed		.60 (Mdn.)
	B	Not Analyzed		.55 (Mdn.)

Notes: 1) A, B, H, & NAP denote Anglo, Black, Hispanic, and Native American Papago, respectively
2) The Metropolitan Achievement Test was used in the Pima County Study.
3) The California Achievement Test was used in the Austin, TX Study.

groups. To return to our original questions at the beginning of this section, the presently available evidence does indicate that the WISC-R is unbiased on the criterion of predictive validity. This conclusion of course must be made somewhat conditional due to some variations in studies and insufficient evidence concerning all groups of potential interest.

BIAS IN TEST USE: SOCIAL CONSEQUENCES

The previous definitions of test bias, although important, are inadequate in terms of the overall influence of tests upon the lives of persons. Testing does have social consequences. Tests, even those which predict accurately, have been misused to justify race, social class, and ethnic discrimination. Kamin (1974) and Cronbach (1975) provided ample evidence regarding the misuse of tests to justify racial and ethnic discrimination in the early decades of this century. What was surprising to me was the rather frequent use of IQ test results to justify racial segregation in public schools during the 1960s. Bersoff's (1979) review of the litigation regarding these practices demonstrated clearly that misuse of test results to justify discrimination was not simply an unfortunate event in

American psychology that occured a long time ago, but that such misuses have occured fairly recently. Further, the implications for and occasional recommendations regarding social policy that are justified today by citing group differences on IQ tests are potentially as discriminatory and abusive with regard to individual and group rights as anything done in the past (e.g., Schockley, 1971). IQ test results have sometimes led to a reduction of opportunities for persons and have qualified persons for apparently ineffective interventions which may have been stigmatizing and humiliating. At the same time, we would be remiss if we didn't emphasize that standardized test results have also been instrumental in removing existing barriers and in increasing opportunities for many minority persons. However, to defend tests simply on the basis of predictive accuracy is to miss entirely the points raised by recent critics of tests.

Jackson's (1975) response to the report of the American Psychological Association Committee on Educational Uses of Tests (Cleary, Humphreys, Kendrick, and Wesman, 1975) is even more to the point. Jackson saw the report as largely irrelevant to the concerns expressed by minorities. The report defended the technical adequacy of the tests when in fact the major concerns of Black and Chicano psychologists (Bernal, 1975) are with how tests affect the lives of persons. The fact that tests have been used by some to justify racist ideology, and otherwise have been misused or misinterpreted in inferences about the potential of individuals are facts acknowledged even by the authors of the APA report. Thus, to defend tests on the basis of evidence of common regression systems, or to attempt to separate the issues of technical adequacy from those of social consequences is insufficient for our purposes of attempting to enhance fairness in test use.

Overinterpretation, Misinterpretation, and Misuse of Test Results.

Much of Mercer's recent work would appear to be directed quite properly toward eliminating the misinterpretation of IQ test results. The issues that have become involved in the debate over SOMPA have occasionally led the discussion away from this very crucial effort. Mercer emphasizes that all current ability and aptitude tests are measures of learning. There should be no disagreement over this point if we merely consider the content of tests and the constitutional repertoires of human infants. It is true that many, perhaps most skills measured by test items do depend on certain maturational developments, but learning after the maturational readiness is achieved is still necessary for mastery of the skills. Therefore, in a general sense IQ tests such as the WISC-R clearly are tests of learning.

It is not difficult to locate numerous examples of overinterpretation of the WISC-R. For example, use of the WISC-R subtest patterns or differences between Verbal and Performance Scale IQs as the basis for a diagnosis of learning disabled, mild mental retardation or even emotional disturbance is all too common. These diagnostic inferences are part of longstanding tradition (and folklore) in applied areas of psychology. Certain technical problems such as unreliability of difference scores and the dangers of making generalizations to individuals from studies of intact groups have been known, but not appreciated sufficiently for many years. Recent data on the base rates of subtest fluctuations and IQ scale differences should certainly reduce this sort of overinterpretation of the WISC-R. Kaufman (1979a;b) reviewed data from the WISC-R standardization sample which demonstrated unequivocally that subtest fluctuations and IQ scale differences are the rule, not the exception for normal children. Continued use of the WISC-R patterns to establish or even support a differential diagnosis is clearly indefensible. Readers interested in these data are referred to Kaufman's very clear discussions of appropriate interpretation of the WISC-R.

Unfortunately, the kind of overinterpretation described in the preceding paragraph probably is not the most serious misuse of IQ test results. Results from intel-

ligence tests such as the WISC-R are all too often believed to be fixed, unitary, and predetermined by genetic factors. These myths are too prevalent among consumers of test results, e.g., parents and teachers, and even perhaps among school psychologists, for us to ignore. Reactions to these myths which lead to misinterpretation and misuse of intelligence test results are among the most frequent concerns expressed by critics of intellectual assessment with minorities. These myths were also a major underlying concern in the placement litigation of the early 1970s (Reschly, 1979).

The myths that IQ test results are fixed and that intelligence is unitary are relatively easy to refute. I know of no one in the field who argues that present IQ tests measure all or even a majority of the important capabilities and competencies related to success and overall adaptation. Certainly the authors of major tests such as David Wechsler recognize that even our very best instruments do not measure everything of importance, and that intelligence is a many-faceted, not a unitary, attribute of the individual. The fact that IQ scores are not fixed, i.e., do not stay constant, is readily apparent from careful examination of data from longitudinal studies (McCall, Appelbaum, and Hogarty, 1973). It is true that scores on IQ tests are fairly stable after age 6 for groups of individuals. However, the IQ scores for a significant percentage of individuals (at least 20 percent) change by 15 points or more between age 6 and maturity, and considerably larger changes of 30 or 40 points have been reported for a few cases. When large changes do occur they tend to be associated with significant changes in the individual's environment or overall emotional adjustment. The fact that IQ tests do change as a function of changes in the individual or the environment might be seen as evidence for increasing our confidence in the test results as indicators of current intellectual functioning, probably the most common interpretation of IQ test results. We need to be conscious of the fact and inform others that scores do change, and that inferences about the future intellectual status of the individual are always tentative.

The final myth, that IQ is predetermined by genetic factors, is a bit more complex. As noted earlier the information or problem solving skills required on IQ test items are learned. However, this fact does not preclude the influence of genetic factors on test scores. Although nearly irrefutable data exist to confirm that genetic factors influence measured intelligence, the unanswered (and unanswerable) issues are the amount of influence attributable to genetics and the genetic influence on the score for an individual. Discussion of the first "unanswerable" question is far beyond the scope of this chapter. Consideration of the question of the genetic influence on the score of an individual is a central issue in resolution of problems of misinterpretation of IQ test scores.

Mercer (1979) provided an excellent summary of the precise conditions that must be met in order to legitimately interpret the differences in scores of individuals (or groups) as reflecting different levels of innate potential. These conditions are: 1) Equal exposure to opportunities to learn the information or problem solving skills measured by the test; 2) Equal levels of motivation to learn and reinforcement for learning whatever the test requires; 3) Equal familiarity with tests and test-taking situations; 4) Persons (or groups) being compared are equal on affective factors such as anxiety, fear, and emotional turmoil which might interfere with learning or performance on the test; 5) Persons (or groups) being compared are equal on physical, sensory, or motor abilities which might interfere with test performance or with learning. Meeting these criteria in any practical situation in which the WISC-R is part of the assessment battery for an individual is virtually impossible. I might add that meeting the criteria or controlling their effects in research on groups is very rarely, if ever, possible.

Those of us who conduct intellectual assessments with tests such as the WISC-R have a special responsibility to protect our clients from misinterpretation of test results. Several courses of action appear to be needed at the present. The myth that intelligence is unitary, and the only attribute of a student that we consider to be important in classification and programming, can be dispelled most effectively by carrying out the full multifactored assessment requirements of PL 94-142 (Tucker, 1977; Reschly, 1979). Many of the PL 94-142 requirements, particularly the phrases "No single procedure is used as the sole criterion..." and "...to assess specific areas of educational need and not merely those which are designed to provide a single general intelligence quotient:" appear to be designed to alleviate past misuses of intelligence test data (Federal Register, 1977). We can argue about, and I believe refute, the notion that IQ tests were used as the single source of information in previous classification and placement decisions involving minorities. However, the documentation provided for classification and placement decisions often appeared to place primary, if not sole, reliance on IQ test data. Implementing and documenting the multifactored assessment requirements should both dispel any remaining misconceptions that we believe intelligence to be unitary, as well as lead to better classification and programming decisions.

Another desirable step in reducing misconceptions would be to change the name of the construct that IQ tests measure. The validity evidence for IQ tests indicates relatively strong predictive validity for performance in academic settings. This relationship is certainly not trivial, and can be shown to be related to other variables such as occupational attainment (Matarazzo, 1972). However, the relationship is somewhat limited. In recent work I have suggested the term "academic aptitude" as a more accurate characterization of what the WISC-R and other IQ tests actually measure (Reschly, 1979). Mercer (1979) suggested the term "School Functioning Level" (SFL) which appears to be motivated by the same concern regarding reducing misinterpretation of IQ test results. Changing the name is of course not a panacea for misinterpretation. It is a step in that direction.

In view of the continuing problems with misinterpretation of IQ test results by consumers of test information, particularly parents and teachers, we developed the following statement for use in school psychology practicum work at Iowa State University. We believe the statement might be used as a kind of "Surgeon General's Warning" about IQ that should appear on reports, protocols, and perhaps, in test manuals. It is consistent with our belief that misunderstanding IQ test information could be damaging to the "psychological health" of the child.

> IQ tests measure only a portion of the competencies involved with human intelligence. The IQ results are best seen as predicting performance in school, and reflecting the degree to which children have mastered middle class cultural symbols and values. This is useful information, but it is also limited. Further cautions-IQ tests do not measure innate-genetic capacity and the scores are not fixed. Some persons do exhibit significant increases or decreases in their measured IQ.

I'm sure the statement could be improved. Perhaps the task of developing an appropriate statement should be referred to the committees in NASP and APA Division 16 that deal with social issues which I believe this certainly is. In any event, it reflects our desire to reduce misinterpretation of IQ tests, which yield information we consider and encourage others to consider as valuable, but limited.

Selection Ratios.

One of the most important social consequences of the use of IQ test information in classification and programming decisions is that disproportionate numbers of certain minorities may be deemed eligible for special education programming. Overrepresentation of minorities in special education programs was the initial complaint in the placement litigation of the early 1970s where the courts implicitly used the rather simple notion of selection ratios as evidence of bias (Reschly, 1979). While we are considering the issue of overrepresentation, some clarification of the percentages cited to establish the disproportionality is in order. In the Larry P. case, indisputable facts were that Black students constituted about 10% of the total student enrollment in the California public schools, and that about 25% of the enrollment in special classes for the mildly retarded was Black. I suspect that many have made the totally erroneous conclusion that many if not most Black students were in programs for the mildly retarded. The analysis of California enrollment data in Table 5 indicates that only a small percentage of Black students were placed in special classes for the mildly retarded. These data certainly do not support the extreme criticism that the primary purpose of IQ tests is to label minority children as "uneducable".

Table 5

Analysis of California Enrollment Data

	1968-69	1976-77
Total student enrollment	4,500,000	4,380,000
Total black enrollment (10%)	450,000	438,000
Total enrollment in special EMR classes	57,148	19,289
Black enrollment in special EMR classes	14,573	4,899
	(25.5%)	(25.4%)

Percent of total student enrollment placed in special EMR classes.

1968-69: 57,148 4,500,000 = 1.3%
1976-77: 19,289 4,380,000 = 0.4%

Percent of black children placed in special EMR classes.

1968-69: 14,573 450,000 = 3.2%
1976-77: 4,899 438,000 - 1.2%

The precise role of IQ tests, most often the WISC-R, in the referral, assessment, classification, and placement process with minorities is not entirely clear. The courts seemed to assume that IQ tests were the primary factor in this entire process, and thereby the major cause of overrepresentation of minorities. This assumption is probably an oversimplification of the actual course of events. Meyers, Sundstrom, and Yoshida (1974) pointed out that IQ testing follows teacher referral and therefore is not the first nor perhaps, even the primary step in the process. Mercer (1973) reported that some children with IQs below the eligibility cut off scores are never referred (and therefore not assessed or placed) while some others with IQs above the cut off scores are referred and assessed, but not placed. This raises an intriguing question. What has been the overall effect of the WISC-R on proportions of minorities classified and placed? Is the effect of IQ testing to increase or decrease the overrepresentation that would occur if the primary criteria for placement were classroom grades and teacher referral? Although the data on this issue are quite limited, there is some evidence indicating the overall effect of IQ tests is to protect minorities from misclassification (Ashurst and Meyers, 1973). More data on proportions of children from diverse groups who fail on various criteria

at the different stages of grades, referral, formal assessment, classification, and placement would clarify this issue. Although there is much opinion to the contrary, it appears likely that IQ tests have served to reduce, not increase the proportions of minorities classified and placed in special education programs.

Application of the multifactored assessment requirements may reduce the overrepresentation of minorities in the future though this is not clear at the present. The overall effect of using a broader variety of information on classification and programming with minorities will likely be determined by how adaptive behavior and sociocultural background are conceptualized, measured, and used. If adaptive behavior is conceptualized narrowly as nonacademic social role performance only, measured with instruments such as the SOMPA Adaptive Behavior Inventory for Children, (ABIC) with a low score required for classification and placement, then overrepresentation is likely to be reduced, perhaps substantially (see later section). Use of sociocultural information to reinterpret WISC-R scores as in SOMPA might also have the effect of reducing the overrepresentation. We can only speculate on the question of whether these changes would be beneficial to minorities.

Summary

Is the WISC-R biased against minorities? Is the WISC-R valid when used with minority children? This section has been devoted to a discussion of these seemingly simple questions. However, the answers are complex and tentative. Decisive and unequivocal conclusions are impossible due to the diverse conceptualizations of the basic problem of bias and the somewhat limited data base.

Conclusions regarding validity and bias of the WISC-R with minorities obviously vary depending on the definition of bias. As noted earlier, there is no single "correct" definition of bias. If definitions are used which stress various interval and external criteria, the research evidence suggests the WISC-R is both valid and unbiased when used with minorities. Other definitions which stress mean differences, selection ratios, and the social consequences of test use result in the opposite conclusion, i.e., that the WISC-R is biased and depending on the value judgments applied to specific situations, perhaps invalid as well.

The reassuring evidence regarding the internal and external validity of the WISC-R provides a foundation for our efforts to eliminate the other possible sources of bias. Of particular concern is the evidence that the results from intelligence tests such as the WISC-R have sometimes been misused to justify race, class, and ethnic discrimination; have sometimes been misinterpreted as indicating innate potential; and have been part of a process whereby minority students were placed in programs that allegedly were ineffective. These undesirable social consequences of test use, although not universal and not an intrinsic characteristic of the test, have been too common for us to ignore. Elimination of discrimination, correct interpretation, and effective interventions are essential components of the effort to ensure useful and fair assessment for all persons. The WISC-R can be a valuable instrument in that effort.

OUTCOMES CRITERION

The most damaging allegation by minority critics of intelligence (academic aptitude) tests is that through their use minority children have been differentially exposed to ineffective educational programs which also had the effects of creating stigmas, reducing self-concept, and restricting career opportunities. Based on the review of the WISC-R to this point, it would appear that the fundamental problem is the outcome of test use, not the test per se. This allegation, however, is serious.

Failure to understand this concern has probably contributed to the poor communication between critics and proponents of tests like the WISC-R. We (proponents) have focused on various internal and external criteria of validity while the critics have raised the broader, and clearly legitimate, question of what happens to minority children as a result of test use.

One result of test use with minorities has been overrepresentation in special education programs. Are these programs effective? The evidence to date, although enormously complex, is not particularly positive at least for the special class kind of intervention. It should be noted that this evidence is the subject of considerable debate (Kolstoe, 1976). However, if the special class programs are as ineffective as some critics charge (e.g., Dunn, 1968), then no child regardless of ethnic or racial status should be placed in the programs

In an effort to focus attention on the what was conceived as the overridding issue in test bias, i.e., the outcomes of test use for the individual, the following definition of bias in assessment was developed.

> Assessment which does not result in effective interventions should be regarded as useless, and biased or unfair as well, if ethnic or racial minorities are differentially exposed to ineffective programs as a result of assessment activities (Reschly, 1979).

The two essential components of this definition of test bias are usefulness and fairness. Usefulness in the sense of assessment resulting in effective interventions that improve skills and competencies, and thereby enhance opportunities, is a paramount goal of school psychological and special education services. The usefulness of assessment instruments such as the WISC-R should be determined on the basis of the degree to which they contribute to realization of this goal. It is acknowledged that there are some instances in which assessment leads to accurate diagnoses for which there are no known effective interventions. These diagnoses may still be "valid" in the sense of validity used by Cromwell, Blashfield, and Strauss (1975) if they improve estimations of prognosis or contribute to prevention of the condition in future cases. However, accurate prognostic estimates or prevention of the condition in future cases are rarely of benefit to the individual being assessed if effective interventions cannot be developed.

In this conception of bias in assessment the concern for fairness is closely related to the notion of usefulness. Assessment and accompanying diagnoses are seen as biased or unfair if they result in overrepresentation of minorities in programs that are ineffective, or in no planned interventions at all. Under such circumstances, the diagnosis may be accurate and the assessment conducted competently, but it is difficult to identify any benefit to the individual. Moreover, if there is a negative connotation or stigma associated with a diagnosis which occurs more often with individuals from minorities, the assessment leading to that diagnosis would be regarded as biased or unfair in the above circumstances. On the other hand, assessment which leads to accurate description of current behaviors, to diagnoses which are essentially summary statements of these behaviors, and to effective interventions, should be regarded as fair or unbiased regardless of the ethnic or racial composition of student groups. Over or underrepresentation of minorities in various classifications or programs is therefore not sufficient to establish bias from this conception.

A number of factors can be identified as prerequisites to achieving fairness in assessment using this approach (Reschly, 1979). However, the more narrow test based criteria discussed earlier in this chapter are usually necessary conditions for fair-

ness in assessment. In order to implement this conception of nonbiased assessment, the tests used must be reliable and valid for all groups; the test results must not be unduly affected by situational-examiner effects; and the content of the tests must reflect important domains of behavior for all groups. These conditions are all necessary, but not sufficient conditions for assessment to be useful and unbiased.

PREREQUISITES TO NONBIASED ASSESSMENT

In this paper, nonbiased assessment is defined in terms of outcomes for the individual. Assessment that is useful in relation to providing effective educational and psychological interventions is regarded as fair, and beneficial to the individual. Valid and reliable assessment instruments are necessary conditions to achieve this goals. Other variables such as what to assess, the link between assessment and programming, effective alternatives, etc. also are necessary conditions. The broader context for this discussion is good fundamentals in assessment and ethical professional practices.

Good Fundamentals and Ethical Practices

This is not the proper forum for an attempt to specify all the competencies needed by related services personnel, or the major provisions of professional ethics. However, these areas are crucial to fair and useful assessment. In some of the placement bias cases there were well documented instances of outright incompetence and clearly unethical practices. Although these cases are probably rare, they do establish the need for all of us to assume direct responsibility for the quality of our services, and indirect responsibility for the professional work of our colleagues.

Clarification of Purpose

Clarification of the purpose for assessment activities is an important, but frequently ignored aspect of good fundamentals. Salvia & Ysseldyke (1978) provide an excellent description of the usual purposes for assessment in remedial and special education. Related services personnel such as psychologists and social workers typically engage in assessment activities for two purposes; Classification/Placement or Program Planning/ Intervention. These two purposes usually involve different types of decisions and different types of instruments.

The Classification/Placement purpose typically involves decisions about current level of performance, degree of discrepancy from grade or age expectancies, degree and type of need, and eligibility for special programming. The questions typically are addressed from the perspective of a comparison of the individual student's performance in relation to some group, usually a representative sample of other students. In recent years these comparisons have been called norm referenced.

Assessment instruments and other data collection procedures for classification/placement decisions should meet certain requirements. The items should be representative of some domain of behavior. The sample of items (or observations) should be sufficient to infer the individual's level of competence in the area. The inferences about degree of discrepancy from expectations should be based on comparisons to a representative sample, i.e., good norms. The scores used in these comparisons should have relatively equal units throughout the scale, and so on. The scores should be highly reliable if decisions are made about individuals. If the scores for a particular instrument are not highly reliable (e.g., .9 or above) then multiple sources of information using different instruments or data collection procedures should be devel-

oped and considered in making decisions. Finally, if inferences are made about underlying traits such as intelligence or psychological processes, the instrument must have good predictive validity relative to appropriate critrion behaviors in educational settings.

Program Planning/Intervention decisions require somewhat different types of assessment information and different types of instruments. Rather than general degree of need or overall strengths and weaknesses, information is needed on very specific skills or competencies. Data collection from this perspective, often called criterion referenced now, is designed to pinpoint precisely what the child can and cannot do in some important domain of behavior. The items on such instruments should provide thorough coverage of the important skills or competencies rather than representative sampling. The items or observations should be related to important objectives and, ideally, to clearly specified interventions.

Most current instruments or observation procedures do not meet the necessary criteria for both purposes. In nearly all cases, a particular instrument or observation procedure has desirable characteristics for norm referenced, classification/placement decisions OR criterion referenced, program planning/intervention purposes. Of course, many instruments do not meet the criteria for either. Many of the mistakes in assessment work originate in failure to clarify purpose. Sometimes we attempt to use the same instrument for both purposes, e.g., use of the WISC-R to suggest educational programming objectives and to determine eligibility for special programs. The WISC-R has many desirable features for certain classification/placement decisions. It is largely irrelevant to specific decisions about educational programming.

Clarification of purpose will lead to different and more varied strategies in assessment.

Relevant Assessment

Assessment which meets the outcomes criterion suggested in this paper must be relevant to educational programming, or in the words of the PL 94-142 Rules and Regulations, "...tailored to assess specific areas of educational need...". A number of current trends in assessment practices enhance the relevance of assessment.

Assessment-Intervention-Evaluation. Assessment for classification/placement is important, but insufficient in relation to the outcomes criterion. Related services personnel increasingly have the opportunity to be involved with other types of assessment such as assessment for: 1) Decisions about special education program option, e.g., resource vs special class; 2) Intervention goals; 3) Intervention strageties; and 4) Evaluation of intervention outcomes. In addition, school psychologists and social workers have opportunities to use behavioral consultation strategies in the home and school. These strategies, involving behavioral assessment procedures, reflect one of the clearest examples of the overall link between assessment, intervention, and evaluation of outcome (Bergan, 1977).

The PL 94-142 requirement that a member of the diagnostic team serve on the committee which designs the initial IEP provides the opportunity for most related services personnel to become more involved with decisions about interventions. Many will have opportunities to participate in annual reviews, and nearly all will be involved with the mandated re-evaluations every three years. The three year re-evaluations are often downgraded in the priorities of related services personnel. This is indeed unfortunate. One of the important questions in this re-evaluation is classification or continued eligibility. Perhaps even more important is careful evaluation of

the effectiveness of the special education programming, and examination of the areas of educational need. How we view the re-evaluations will be heavily influenced by whether we see ourselves as classification personnel, OR whether we adopt the outcomes criterion. Nevertheless, the opportunities now exist for significantly greater involvement in all phases of designing, carrying out, and evaluating interventions.

Reduced Level of Inference. Relevant assessment involves a lower level of inference. School and other areas of applied psychology have an unfortunate tradition of combining "clinical insight" with very minimal data resulting in global descriptions of persons. Many of the standard interpretations of test results involve analogical reasoning with little or no empirical support. The analogical reasoning used in the interpretation assumes that a logical relationship exists between the observed behavior and underlying dynamics, or, nothing is quite what it seems to be. Usually the empirical support for the interpretation simply does not exist, or the strength of the relationship, although statistically significant, is so low that prediction for individuals is hazardous at best. An example may clarify these points. A common interpretation of dark, heavy lines on the Bender designs is "repressed hostility" even when the designs are reproduced accurately. This "emotional indicator" is frequently discussed in reports without any additional or external verification even though the empirical evidence is weak (Koppitz, 1975, p. 85). These "signs" may provide cues to important behaviors that should be assessed in relevant situations. However, the sign as such is based largely on analogy, likely to be inaccurate for the individual, and even worse, may impede efforts to develop interventions. Similar reasoning and interpretations for a variety of other tests are found in standard clinical texts (e.g., Rapaport, Gill, & Schaefer, 1968), which are frequently used in school psychology training.

Another change related to the reduced level of inference is less emphasis on underlying dynamics. The frequent question at staffings after potentially useful objective information is presented is "What is really going on?" This question often serves as a cue for all manner of speculation about "pathological" family dynamics, who perceives whom as what, juicy anecdotes about sexual proclivities, and so on. These speculations, and the high level of inference upon which they are based, might be useful IF effective interventions were the result. The usual outcome, however, is participant satisfaction over their apparent insight and understanding regarding the problem. These underlying dynamics are rarely used to design interventions if for no other reason than the impossibility of influencing the variables involved. If the question of "What is really going on?" leads only to speculation without specific interventions, then the entire exercise should be regarded as professional voyeurism. At a minimum, it is useless assessment.

There are several trends which will continue to move the field toward a reduced level of inference and less emphasis on underlying dynamics. One influence is the courts as well as the quasi-legal appeal procedures established as part of the due process regulations. Speculative inferences based on minimal data have not been well received by the courts (Ziskin, 1975). Another influence is the PL 94-142 requirement that tests be validated for the purposes for which they are used. Presumably, testimonial evidence from satisfied clinicians will not suffice. Finally, the strong emphasis on designing interventions, and on review and evaluation of interventions will necessitate greater consideration of other more useful information.

Situational or Behavioral Assessment. Behavioral or situational assessment is perhaps the most rapidly expanding model of assessment today. The behavioral approach with the emphasis on precise formulation of goals, careful observation of situational factors, implementation of specific interventions, and evaluation of outcomes is consistent with

many requirements of recent legislation and the outcomes criterion proposed in this paper.

The behavioral approach has been regarded as too restrictive by many school psychologists. Somehow the emphasis of behaviorists on functional control rather than explanation and understanding has appeared to reduce school psychology to technology rather than "science" or profession. Those who have made these judgments in the past are encouraged to reconsider the question of theoretical model through reviewing the advances of the past decade in behavioral theory, assessment, and interventions (Bergan, 1977; Cone & Hawkins, 1977; Keller, 1980; Michenbaum, 1977). Attention also is directed to a recent article suggesting a behavioral perspective on the use and interpretation of intelligence tests (Nelson, 1980). Behavioral models now include and operationalize broad classes of behavior such as cognitive style, social skills, anxiety, etc. The behavioral assessment techniques have been refined to include a broad variety of instruments and observation methods for collecting useful information, many of which are relatively unobtrusive in natural settings. Perhaps the greatest advances have been in the use of more natural interventions such as self-control, cognitive self-instruction, rehearsal, modeling, and naturally occurring reinforcement contingencies.

Placement Options and Effective Programs

If we accept the notion that possible bias in assessment is best conceptualized in terms of outcomes, then the availability of effective educational programs and alternative placement options is an absolute prerequisite to implementing nonbiased assessment procedures. In the situations which resulted in the special education placement litigation, the educational programs were presumed to be ineffective and the range of options limited. The author remembers all too well the very limited range of options that was typical until quite recently. The only choices often were regular classrooms with no assistance or self-contained, segregated classes for the mildly retarded. Many psychologists can recall vividly cases where we knew the child was not "really" retarded, but in view of very low achievement accompanied by increasingly negative attitudes toward school and self, the self-contained, segregated class appeared to be the best option.

This situation has changed, or is in the process of change. A wide range of options are increasingly available, the principle of using the least restrictive alternative is the law of the land, and greater emphasis is placed on effectiveness of interventions through individualized educational programs with annual reivew. These changes provide the opportunity for assessment activities in a broader variety of areas. In addition to classification decisions, assessment should be directed toward decisions concerning choice of least restrictive alternative and toward the content of interventions, especially identifying specific areas of "educational" need in terms of social, emotional, and academic development. Assessment should also yield information concerning the approach to intervention, specifically, changes in antecedent, situational, and consequent environments that can be used to carry out interventions. Finally, we need to gather information that is relevant to and/or can assist others in evaluating the effectiveness of interventions.

Multifactored Assessment

The concept of multifactored assessment apparently was the primary solution to the dilemma of defining and describing the requirement of nonbiased assessment in the PL 94-142 Rules and Regulations. The underlying (and logical) assumption is that assessment is likely to be less biased if a broad variety of information is collected and considered systematically in making classification/placement decisions. This assumption is sound, but insufficient. Improved classification decisions are certainly

important, but even more important is the use of the multifactored information in designing and evaluating interventions.

Tucker (1977) provided a description of the categories of information which should be developed in a comprehensive assessment of children "for possible mildly handicapping conditions." For the most part, the categories of information are fairly standard and largely consistent with traditional descriptions of comprehensive psychoeducational evaluations. The arrangement of the categories of information, especially the sequence suggested for collecting the information, is somewhat unique. These categories have been further modified through the concepts of low and high inference procedures in the scheme presented in Table 6. It should be noted that several activities should occur before the preplacement evaluation is initiated (See Guidelines at the end of this paper). Among these activities are screening of referrals, clarification of referral problem(s), interventions within regular education, etc. If these procedures are followed, i.e., it is determined that a severe discrepancy exists and regular education alternatives have been unsuccessful, then the preplacement evaluation should be initiated.

Table 6

Multifactored Assessment

A. SCREENING PHASE

1. Referral. Clarify referral through teacher interview, classroom observation, and examination daily work.

2. Educational history. Review current and previous educational records including special services, classroom performance, standardized tests, etc. Consider use of regular education options and interventions.

IF THERE IS A SEVERE DISCREPANCY, OR IF THE DEFICIT IN PERFORMANCE IS COMPREHENSIVE AND LONG TERM, AND IF REGULAR EDUCATION OPTIONS HAVE BEEN ATTEMPTED UNSUCCESSFULLY, THEN, INITIATE THE PREPLACEMENT EVALUATION.

B. PREPLACEMENT EVALUATION (Initial Phase)

3. Procedural Safeguards. Follow procedural safeguards to meet legal requirements and to establish communication with home.

4. Multidisciplinary Team. Form multidisciplinary team, develop hypotheses, tailor the preplacement evaluation to the individual, assign responsibilities and establish time lines.

C. PREPLACEMENT EVALUATION (Low Inference)

5. Sensory Screening, Health, Developmental. If needed, physical examination (if needed), health and developmental history, and sensory assessment by specialists.

6. Language Dominance. Determine the child's primary language competence through formal measures and/or home interview.

7. Educational Evaluation. Determine level, pattern, strengths and weaknesses in academic skills through formal and informal measures administered and interpreted by specialists.

D. PREPLACEMENT EVALUATION (High Inference)

8. Perceptual-Motor/Psychological Process. Determine if severe process deficits are related to learning problem through administration of formal instruments, observation, and interview.

9. Adaptive Behavior-Outside School. Investigate social competence outside of school through structured and unstructured interview.

10. Social/Emotional. Determine nature and extent of social/emotional involvement or behavior disorders through interviews, observation, checklist, etc.

11. Intelligence (Academic Aptitude). Determine general level of expectations for academic achievement through administration of individual intelligence test.

E. DECISION-MAKING

(See Guidelines at end of paper)
Consider this criterion: Would you be satisfied IF YOUR CHILD HAD BEEN INVOLVED IN THIS ASSESSMENT PROCESS?

There is nothing new about the concept of a multifactored assessment. Professional Standards have always emphasized the importance of collection and consideration of a broad variety of information as a part of any significant classification/placement decision. Implementation of this notion has been less consistent. Even more troublesome, documentation through reports and other records of the full multifactored process has not been universal. For example, in presenting a comprehensive record for a child classified and placed in special education, it is important to thoroughly describe the initial referral and educational history, not just the intelligence test data. In the past, the records for students in special education programs often had little information beyond the intelligence test results. Other types of information probably were collected and considered in most cases, but were not documented.

The recent versions of the multifactored assessment reflect greater emphasis on sources of information other than intelligence test data. This suggests the very proper concern that intelligence not be the sole or primary source of information for classification/placement decisions. Moreover, the information collected as part of the low inference procedures described in Table 6 will, in some cases, significantly influence the selection, administration, and interpretation of high inference procedures such as intelligence tests. For example, some among us (related services personnel) have had the embarrassing experience of administering a verbal scale to a hearing impaired child, or a performance scale to a child who needed (and had) glasses, but wasn't wearing them that day. These kinds of errors are humorous if corrected, but potentially tragic if allowed to stand. The point is that the low inference procedures should always be conducted before the high inference procedures.

Recent versions of multifactored assessment reflect more emphasis on the three areas of adaptive behavior, primary language, and sociocultural background. These are not totally new areas of assessment. However, the implicit and sometimes explicit requitements that they be assessed systematically and considered carefully create difficult challenges for related services professionals. In subsequent sections, the conceptual and technological bases for these areas will be reviewed.

THE SYSTEM OF MULTICULTURAL PLURALISTIC ASSESSMENT

Discussions of bias in assessment are incomplete without consideration of the System of Multicultural Pluralistic Assessment (SOMPA) (Mercer, 1979). The SOMPA models and measures are particularly relevant to discussions of adaptive behavior and sociocultural background. Much of the rationale for SOMPA is based on an epidemiological study of mental retardation in Riverside, California.

Mercer's Riverside Studies

At about the time that national concern was increasing over the six hour retarded child (see later section), a sociological analysis of the process whereby persons were diagnosed as mentally retarded appeared in the literature (Mercer, 1970, 1973). Although the major findings of Mercer's Riverside, California study were no surprise to professional personnel in mental retardation and special education, the conclusions reached by Mercer called for substantial changes in assessment practices. Of particular importance was the call for greatly increased emphasis on adaptive behavior and sociocultural information.

The major findings of the Riverside Study were that public schools were by a large margin the community agency most likely to diagnose persons as mentally retarded. In comparison to other community agencies, the Riverside schools placed more reliance on the results of individual intelligence tests and used a higher IQ cut off score (79 rather than 75 or 70). Persons classified by public schools as mentally retarded were often poor, of minority status, and situationally retarded. Most were regarded as normal by their families and had not been diagnosed as retarded prior to entering the public school. Mercer attributed these findings to, particularly the overrepresentation of minorities, the use of a higher cut off score by the schools, the failure of the schools to assess adaptive behavior, and the biases in the IQ tests.

The findings reported by Mercer which apparently have been influential in the litigation and legislation came as no surprise to persons familiar with the literature on mild mental retardation. For example Heber commented in the 1961 AAMD Manual, "Impairments in learning are usually most manifest in the school situation, and, if mild in degree, may not even become apparent until the child enters school" (Heber, 1962, p. 73). Further, Farber (1968) reviewed prevalence studies in mental retardation and reported higher rates for mild mental retardation among the economically disadvantaged, and a peak prevalence at the ages of about 10 to 14. Mild mental retardation, in contrast to the more severe levels of mental retardation, has been known for decades to be more prevalent among the poor and economically disadvantaged minorities; to be more common during the school age years; to be largely situational or school related; and impermanent.

Mercer's analysis of the Riverside studies also reflects some misconceptions about the complex process in the public schools whereby children are classified as mildly retarded. The actual role of standardized tests are clearly exaggerated. There are few if any instances where cases of mild mental retardation are sought through group standardized tests of ability or achievement. The use of group ability tests appears to have declined in recent years, and in any event, the results of such tests have never been a significant factor in the classification of students as mentally retarded.

The most significant step in the process whereby students are classified as mildly retarded is teacher referral due to poor performance in the classroom. Related services personnel do not go out to schools and attempt to catch unwitting

victims with their psychometric nets. The only children to whom individual intelligence tests are administered are those who have been referred. Mercer gave slight attention to the importance of teacher referral. She reported that 72% of the students classified as mentally retarded by the schools had repeated one or more years prior to classification. The grade retention data suggests that the problems experienced with the classroom situation were chronic rather than temporary, and that at least some minimal alternatives were attempted within the regular classroom.

Mercer (1973) contended, however, that referral rates were not different among white, black, and Hispanic students. If the referral rates were not different, the clear implication is that what happened after referral was primarily responsible for overrepresentation in the programs for the Educable (Mild) Mentally Retarded. What happened after referral in most cases was of course an individual psychological evaluation by a school psychologist in which an intelligence test was usually administered. However, the referral rate data reported by Mercer included all cases, not just those referred for academic problems. Students referred for possible identification as gifted were lumped together with those referred for academic problems. Data on the racial/ethnic composition of the students referred for academic problems were not provided for the Riverside Study.

Other data sources suggest that significantly more economically disadvantaged and minority students are referred due to academic problems (Tomlinson, Acker, Canter, & Linborg, 1977). The effects of psychological evaluation including intelligence testing on the population of economically disadvantaged minorities referred for learning problems has not been studied adequately. Some data suggest that individual psychological evaluation including intellectual assessment serve to protect minority students from inappropriate classification as mentally retarded. Ashurst and Meyers (1973) reported results from an analysis of all students (N = 269) referred over a three-year period as suspected cases of mental retardation. These data were also from the Riverside, California public schools. Referral rates were considerably higher for minority students. The effects of psychological evaluation were to reduce, not increase, the overrepresentation of minorities that would have resulted from teacher referral. Contrary to the suggestions from Mercer's analysis of the data from Riverside, intelligence test results provided some protection of minority students from erroneous classification.

Although the precise role of intelligence tests in producing overrepresentation of minorities in programs for the mildly retarded continues to be a source of debate, other issues from the debate over the six hour retarded child are equally important. The question of whether economically disadvantaged minorities are overrepresented due to socioeconomic status (SES) or minority status has not been studied sufficiently. Mercer (1973) concluded that SES accounted for some but not all of the overrepresentation of minorities. However, the actual data on mean SES levels of the total EMR population and for the different racial/ethnic groups in the EMR population were not reported. Other concerns to be discussed later involve the conceptions of mild mental retardation and adaptive behavior for school age children.

The conclusions reached in the Riverside Study and from the broader concern for the six hour retarded child have had a profound influence on related services disciplines and special education. Mercer (1973) recommended three major changes in the diagnostic procedures used in the public schools. First, it was suggested that the IQ cut off be lowered to the traditional criterion of about two standard deviations below the mean rather than the higher cut off score used then, and now, in many state education codes. The major justification for this change was, "At this criterion level, persons are least likely to be labeled as retarded who, as

adults, will be able to fill a normal complement of social roles" (Mercer, 1973, p. 221). Implicit in this recommendation is the view that "true" mental retardation is a permanent condition. The second recommendation was that adaptive behavior should be emphasized more in classification decisions. Accompanying this recommendation was a broadening of the conception of adaptive behavior in comparison to the 1961 AAMD Manual (see later section). Finally, pluralistic norms were advocated for the purpose of correcting the bias in IQ tests. The SOMPA represents Mercer's attempts to implement the last two recommendations.

SOMPA Models and Measures

SOMPA is a highly complex and innovative approach that has been the subject of much, sometimes acrimonious, debate (see No.'s 1 & 2 of Vol. 8, School Psychology Digest). I encourage all school psychologists to study this approach carefully, and look forward to research on applications of SOMPA. The unfortunate trend currently is toward extreme reactions, positive and negative, ranging from those who "feel" that it is the best thing that has ever been developed to those who "feel" that SOMPA represents a diabolical plot against school psychologists, special educators, children, and so on. The debate has often been useful, but suspension of judgment until more empirical information is available is clearly indicated. The author's, publisher's, and critic's claims and views notwithstanding, we need much more information before reaching firm conclusions.

At the present, SOMPA provides three major innovations concerning assessment practices. The specification of three models of assessment in terms of assumptions, values, and appropriate instruments is one of the major components as well as a controversial aspect of the system. A second innovation is the development of new instruments such as the Physical Dexterity Battery, Sociocultural Scales, Health History Inventory, and Adaptive Behavior Inventory for Children (ABIC). Many of these instruments will be useful data collection devices regardless of the outcome of the debate on other features of SOMPA. Finally, SOMPA combines the models with conventional and new data collection devices to develop a more refined classification system. It is important to note that the primary information from SOMPA at the present is of a classification, not programming, nature. Techniques to use SOMPA information in completion of the full diagnostic construct criteria (Cromwell, Blashfield, & Strauss, 1975) are at present not available. The ultimate usefulness of SOMPA will be determined by the degree to which the information provided is related to educational placement and programming decisions, a point which the authors of SOMPA have also stressed.

Several specific issues need to be addressed in the near future regarding uses of SOMPA. There is the question of the generalizability of SOMPA normative data to other groups, e.g., Native Americans, and to the same groups in different geographic regions, e.g., Hispanics in the Northeast. The SOMPA standardization data are based on carefully selected samples of children, but sample selection was restricted to California. The population in California, although diverse, is not necessarily typical of samples elsewhere, e.g., Anglos in Iowa; Blacks in rural Alabama, or Hispanics in New York City. The authors of SOMPA suggest collection of data from random samples of children in different localities to determine if the SOMPA California norms and regression formulas are appropriate for specific groups of children. Such studies, although expensive, are clearly necessary prior to widespread use of the system. A second issue is related to the generalizability to other groups of the data on the relationship of the WISC-R to other measures in SOMPA. Finally, there is the issue of the effects on children, particularly in terms of educational classification and programming, of use of SOMPA. Limited data on these questions are now available and will be discussed in later sections.

ADAPTIVE BEHAVIOR

Concern for what is now called adaptive behavior is not new. The term social competence was used prior to about 1960 to refer to approximately the same construct. Social competence or adaptive behavior has also been a fundamental concept throughout the history of efforts to describe and explain the phenomenon of mental retardation.

Although the construct of adaptive behavior is not new, a number of recent events have led to considerably more emphasis on use of adaptive behavior data in special education classification and placement decisions. Revisions of the AAMD Manual on Terminology and Classification in 1961 and 1973 reflected increasingly greater emphasis on adaptive behavior. The "normalization" effort which has the primary purpose of integrating institutionalized mentally retarded persons into community settings was a second major influence on adaptive behavior. From this perspective adaptive behaviors are viewed as the "reversible" features of the more severe levels of mental retardation (Leland, 1978). Another somewhat unrelated trend was the emphasis on nonbiased assessment that resulted from litigation and legislation in the 1970's. Adaptive behavior from this perspective was seen as a means to reduce the emphasis on intelligence test results; to provide more equitable assessment for minorities; and to alleviate the overrepresentation of minorities in special education programs for the mildly retarded (Coulter and Morrow, 1978).

In view of the diverse influences and different purposes underlying the recent upsurge of interest in adaptive behavior, it is not surprising that much confusion exists over the measurement and use of adaptive behavior data. In addition to these sources of confusion the recent federal legislation implies that adaptive behavior data must be considered in all special education placement decisions. Perhaps the best recent source of information on adaptive behavior is a book edited by Coulter and Morrow which is cited earlier. Their discussion of unresolved issues surrounding the adaptive behavior concept, available measures, and possible uses is recommended highly.

Adaptive Behavior and Definitions of Mental Retardation

For approximately two decades the AAMD definition of mental retardation has included the dimensions of intelligence and adaptive behavior. However, the emphasis on adaptive behavior was increased in the 1973 version. The 1961 version described mental retardation as subaverage general intellectual functioning which is associated with impairment in adaptive behavior. The 1973 and 1977 versions placed more emphasis on adaptive behavior by changing "associated" to "existing concurrently." This change toward placing relatively equal emphasis on both of the dimensions of mental retardation along with the subtle changes in the conception of adaptive behavior from 1961 to 1973 versions constitute difficult challenges for diagnostic personnel.

By now it is likely that most educational definitions of mental retardation include both the intelligence and adaptive behavior dimensions. According to a recent survey (Patrick & Reschly, 1980) about two-thirds of the states required assessment of adaptive behavior for one or more of the special education classifications, usually mental retardation. A number of additional states reported efforts to add adaptive behavior to the state definition of mental retardation. However, the majority of states did not have a definition of adaptive behavior and much confusion was reported concerning definition, domains of adaptive behavior, and availability of measures. Although the status of adaptive behavior in special education undoubtedly varies from state to state, the trend is toward more emphasis on this

dimension at least with the mentally retarded.

Conceptions of Adaptive Behavior

One of the most influential definitions and descriptions of adaptive behavior is provided in the AAMD Manual on Terminology and Classification. The AAMD conception and criteria for adaptive behavior during the school age years changed in subtle ways from 1961 to 1973. Consider the following description from the 1961 revision.

> "Adaptive behavior refers primarily to the effectiveness of the individual in adapting to the natural and social demands of his environment. Impaired adaptive behavior may be reflected in: 1) maturation, 2) learning, and/or 3) social adjustment. These three aspects of adaptation are of different importance as qualifying conditions of mental retardation for different age groups."
>
> "Learning ability refers to the facility with which knowledge is acquired as a function of experience. Learning difficulties are usually most manifest in the academic situation and if mild in degree may not even become apparent until the child enters school. Impaired learning ability is, therefore, particularly important as a qualifying condition of mental retardation during the school years."

Quotes from Heber, 1961, p. 3-4.

Using the description of adaptive behavior from the 1961 version one might focus attention entirely on performance in the public school context for school age children. Adaptive behavior for school age children in this version appears to be based at least primarily on academic competence. For school age children this conception might be interpreted as specifying a diagnosis of mental retardation based only on intelligence, classroom academic performance, and results of standardized achievement tests. Other characteristics and behaviors specified in current conceptions of a multi-factored assessment should have been and often were considered in mild mental retardation classification/placement decisions. However, the clear implication in the 1961 revision was that academic performance was the most important index of adaptive behavior for school age children. With considerable justification, one could argue that up to 1973 when the AAMD Manual was revised, diagnostic personnel in the schools were assessing adaptive behavior as conceptualized at that time.

The changes in conception of adaptive behavior for school age children in the 1973 and 1977 revisions of the AAMD Manual are illustrated in the quotes below. As noted previously, the 1973 and 1977 revisions are virtually identical.

> "Adaptive behavior is defined as the effectiveness or degree with which an individual meets the standards of personal independence and social responsibility expected for age and cultural group." Grossman, 1977, p. 11.
>
> "During childhood and early adolescence in:
> 5. Application of basic academic skills in daily life activities
> 6. Application of appropriate reasoning and judgment in mastery the environment
> 7. Social skills (participation in group activities and interpersonal relationships)"

> "The skills required for adaptation during childhood and early adolescence involve complex learning processes. This involves the process by which knowledge is acquired and retained as a function of the experiences of the individual. Difficulties in learning are usually manifested in the academic situation but in evaluation of adaptive behavior, attention should focus not only on the basic academic skills and their use, but also on skills essential to cope with the environment, including concepts of time and money, self-directed behaviors, social responsiveness, and interactive skills."

Quotes from Grossman, 1977, p. 13-14.

The recent revisions of the AAMD Manual placed more emphasis on adaptive behavior AND broadened the concept of adaptive behavior during the school age years. It should be noted that contrary to some recent trends in conceptions and measures of adaptive behavior, the AAMD conception does continue to include performance in academic settings as an important component of adaptive behavior during the school age years. For children in this age group, school performance is a necessary part of the construct of adaptive behavior (see below). However, performance in other social settings should also be considered.

Other conceptions of adaptive behavior have been proposed in recent years (see Coulter & Morrow, 1978; Reschly, 1980, 1981 for reviews). The common features of conceptions of adaptive behavior are emphases on developmental (age appropriate) criteria and consideration of cultural context. Conceptions of adaptive behavior for school age children differ sharply on the issues of: 1) Inclusion or exclusion of the cognitive competencies that underlie adaptive behaviors; 2) The social settings and social roles (school vs out of school) included; and 3) The data source, i.e., third party respondent or direct observation of the individual. In addition, conceptions and measures of adaptive behavior have been developed for different purposes, classification/placement vs program planning intervention, and for different populations, mildly retarded vs more severely retarded.

Assessment of Adaptive Behavior

The purpose of assessment, i.e., the decision that needs to be made about or with a student, is the most basic consideration in the selection of a formal measurement instrument or informal data collection procedure (Salvia & Ysseldyke, 1978). Clarifying the purpose through explicit statements of the decisions to be made is particularly important in the assessment of adaptive behavior.

If the purpose of assessment is program planning/intervention with the moderately, severely, or profoundly retarded, the currently available adaptive behavior instruments are reasonably adequate for most ages. Some instruments have been developed carefully with rigorous measurement and statistical criteria applied to selection of items. A sample list of some of the more prominent instruments is provided in Table 7 which is reprinted from Oakland and Goldwater (1979).

Although a number of adaptive behavior measures are listed in Table 7, it should be noted that only two of them are designed specifically for school age populations of normal, borderline, and mildly retarded persons (the AAMD-School and the ABIC). The primary focus in this paper is with nonbiased assessment which is principally a concern about appropriate classification/placement decisions with mildly handicapped persons. Adaptive behavior is one of the key areas in the multifactored assessment scheme developed by Tucker and mentioned in the PL 94-142

Table 7

Measures of Adaptive Behavior

Measurement scales	Behaviors Assessed: Physical development/sensory-motor/locomotion	Self-direction	Language and communication	Vocational and occupational skills	Economic	Social	Self-help/independent functioning/self-maintenance	With peers	In school	In the family	In the community	Age range	Population type: Clinical	Population type: School	Purpose: Screening	Purpose: Placement	Purpose: Programming	Examiner: Teacher	Examiner: Diagnostician	Examiner: Paraprofessional	Respondent: Teacher	Respondent: Parent/family	Respondent: Child	Reliability and validity of data available: Yes/no	Scores: Grade equivalent/age	Scores: Percentile	Scores: Scaled score	Administration time (minutes)
AAMD Clinical Version (Nihira *et al.*, 1974)	X	X	X	X	X	X	X					3-adult	X			X	X		X	X		X		Yes		X		45-60
AAMD Public School Version (Lambert *et al.*, 1974)	X	X	X	X	X	X	X					7-13		X		X	X	X	X	X	X	X		Yes		X		45-60
Cain-Levine (Cain *et al.*, 1963)		X	X			X	X					5-13				X		X	X	X		X		Yes		X		20
California Preschool (Levine *et al.*, 1969)			X			X						2-5			X			X	X	X	X	X		Yes		X		20
Camelot (Foster, 1974) Behavioral Checklist	X	X	X	X	X	X	X					2-adult				X	X	X	X	X*	X	X		Yes				60
Adaptive Behavior Inventory for Children (Mercer & Lewis, 1978)				X	X	X	X	X	X	X	X	5-11		X		X			X	X*		X		Yes		X	X	60
Preschool Attainment Record (Doll, 1966)	X	X	X			X	X					birth-7				X		X	X	X	X	X		No	X		X	20
Vineland Scale (Doll, 1965)	X	X	X	X	X	X	X					birth-25	X			X			X		X	X	(*)	Yes				20

*With extensive training in interviewing

Reprinted from Oakland & Goldwater, 1979, p. 147.

Rules and Regulations. However, the present level of technology with respect to assessment of adaptive behavior with the mildly handicapped including the mildly retarded is characterized well in the following quotes.

> "The inclusion of adaptive behavior in nonbiased assessment by the use of tests or scales to facilitate comparison of a child with his/her peers is not yet perfected." (CORRC, p. 20, Undated report distributed in 1979).
>
> "Presently, the assessment of adaptive behavior through clinical interviews and observations of the child's behavior in other social systems represents the major alternatives for pupil appraisal professionals, if the goal of assessment is primarily placement. Until psychometric technology provides a variety of suitable and more objective behavior measures, the more informal, and thereby subjective, methods will remain in wide use." (CORCC, p. 21, see above).

Problems with assessment of adaptive behavior also were mentioned prominently in the AAMD Manual. Grossman (1977, p. 20-21) emphasized the following problems:

1) the frequent discrepancies in level of adaptive behavior and level of intelligence with the mildly retarded; 2) the unavailability of adaptive behavior instruments that are suffuciently precise to establish a definite cut off score such as minus two standard deviations from the population mean; and 3) the major limitations with most available instruments such as poor norms and item content selected from studies of institutional populations. In view of these limitations, Grossman suggested that assessment of adaptive behavior must involve a large degree of clinical judgment.

Clearly, the available technology leaves much to be desired with respect to assessment of adaptive behavior with normal and mildly retarded children. A considerable amount of additional work on instrument development and research is needed. However, the picture suggested in the quotations above may be a bit too negative. There has been some instrument development and research in recent years that should be applied to the assessment of adaptive behavior in classification/placement decisions. Judicious use of the results from these instruments along with informal sources of data on adaptive behavior should become a part of comprehensive evaluation that is conducted prior to classification/placement decisions.

Review of Adaptive Behavior Measures for the Mildly Retarded

AAMD Adaptive Behavior Scale - Public School (ABS-PS). The most important influences leading to the development of the ABS-PS were legal requirements in California regarding the classification/placement of students in EMR programs. Other purposes such as providing information for educational programs and remediation were also cited by the authors (Lambert, Windmiller, Cole, & Figueroa, 1975).

The items on the ABS-PS are a subset of items from the AAMD Adaptive Behavior Scale - Clinical (ABS-C). The ABS-C was developed from extensive studies of deficit behaviors among institutionalized mentally retarded persons. The purpose of the ABS-C was to pinpoint behaviors which prevented placement of severely retarded persons in community settings. Once these behaviors are identified, the focus is then on remediation, and eventually, placement in less restrictive settings. The critical point is that the items on the ABS-C were selected from studies of severely retarded persons for the purpose of improving program planning/intervention. The content of the Public School version is the same as the Clinical version except for the deletion of 15 of the original 110 items which were judged to be inappropriate for public school students.

The ABS - Public School is divided into two major sections. The first part might be termed adaptive behaviors since high scores on this section indicate higher social functioning. The second part might be called maladaptive behaviors since the higher the score, the lower the level of social functioning. The nine domains involving 56 items on the first part are Independent Functioning, Physical Development, Economic Activity, Language Development, Numbers and Time, Vocational Activity, Self-Direction, Responsibility, and Socialization. A sample item from the Shopping Skills area of the Economic Activity Domain is:

30. Errands (Circle only one)

Goes to several shops and specifies different items	4
Goes to one shop and specifies one item	3
Goes on errands for simple purchasing without a note	2
Goes on errands for simple purchasing with a note	1
Cannot be sent on errands	0

The second part comprised of 39 maladaptive behavior items has the twelve domains of Violent and Destructive Behavior, Antisocial Behavior, Rebellious Behavior, Untrustworthy Behavior, Withdrawal, Stereotyped Behavior and Odd Mannerisms, Inappropriate Interpersonal Manners, Unacceptable Vocal Habits, Unacceptable or Eccentric Habits, Hyperactive Tendencies, Psychological Disturbances, and Use of Medications. Item 32 of the Hyperactive Tendencies Domain is as follows:

32. Has Hyperactive Tendencies

	Occasionally	Frequently
Talks excessively	1	2
Will not sit still for any length of time	1	2
Constantly runs or jumps around the room or hall	1	2
Moves or fidgets constantly	1	2
Other (Specify _____)	1	2

_____ None of the above

Total_____

The child's classroom teacher is the recommended respondent for the ABS-PS. Respondents are allowed to infer or, if necessary, to guess regarding the child's competencies, particularly those which take place outside of school.

The norms for the ABS-PS are based on a sample of 2600 school age children in California. Norms cover the ages of 7-13. Separate norms are provided by class placement (regular vs types of special classes) for Sections I and II of the ABS. In addition, separate norms by ethnicity and sex are provided for Section II.

The interpretation of the ABS-PS is based on comparison of the individual's profile of percentile ranks to modal profiles of children placed in different educational programs. No standard scores are provided for the domain scores, and no overall score for the major sections is available.

Although the ABS-PS has many limitations, it can be a useful adjunct to clinical judgment in classification/placement decisions, and to a lesser degree, in program planning/intervention decisions. The ABS-PS appears to be more appropriate for lower functioning children in the EMR range. The major weaknesses of the instrument are the following: First, the content validity of the items is questionable in view of the original purpose of the ABS-Clinical version. The item format requires a considerable degree of inference or even guessing. The respondent is the teacher who usually has little information about social role performance outside of school. Finally, the method of interpretation, comparing profiles, is highly subjective in many cases.

The Adaptive Behavior Inventory for Children (ABIC) was developed with the explicit purpose of improving classification/placement decisions with the mildly retarded (Mercer, 1979). The ABIC reflects a strong social system perspective with emphasis on how the child functions in different settings and different social roles.

The ABIC items were selected on the basis of intensive interviews with mothers of children between the ages of 5 and 11. The item pool of 480 questions was reduced to 252 questions on the basis of a questionnaire study. These 252 items were administered to a standardization sample. Ten items were deleted resulting in 242 items in the final published version. Most items are age graded. A basal and ceiling procedure is used in administration of the ABIC. The domains covered by the ABIC are Family, Community, Peer Group, Nonacademic School, Earner/Consumer, and

Self-Maintenance.

Sample items from each of the ABIC domains are provided below.

Domain	Item
Family	147. When ___ cannot have what he/she wants immediately, how often does he/she get angry and fuss about it? 0 most of the time 1 sometimes, or 2 almost never
Community	142. When visiting relatives or friends outside the neighborhood, does ___ usually 0 goes with an older person 1 go with children his/her own age, or 2 go alone?
Peer Relations	144. How often does ___ meet and play with his/her friends at a special place like a vacant lot, a park, the street, the school bus stop, or a courtyard? 1 sometimes 0 seldom or never, or 2 often
Non Academic/School	132. How often does ___ take his/her school supplies and books to school without being reminded? 1 occasionally 0 seldom, or 2 regularly
Earner/Consumer	140. Does ___ make correct change for a dollar 2 without help 1 only with help, or 0 not at all?
Self-Maintenance	143. Does ___ order food at a restaurant 2 without help 1 with some help, or 0 does someone order for him/her?

The ABIC is administered as a structured interview. The primary caretaker of the child, typically the mother, is the preferred respondent. For each item the mother chooses among three possible responses.

Standard scores with a mean of 50 and standard deviation of 15 are provided for each domain. The average of thse standard scores is used as a composite or global index of adaptive behavior. In addition, three other scores are provided. The Veracity Scale attempts to detect a "fake good" response set. The "No Opportunity" and "Not Allowed" responses are seen as an indication of the amount of restriction place on the child. Finally, the "Don't Know" responses are viewed as an indication of the amount of knowledge the respondent has about the child's activities. If critical values are exceeded on the three ancillary scales, interpretation of the other scores is not recommended.

The ABIC norms are based on a stratified random sample of 2100 children in

California schools between the ages of 5 and 11. Stratification variables included sociocultural group (Anglo, Black, and Hispanic), size of community, and gender.

The ABIC is the only instrument in which the entire design from item selection to standardization was directed toward classification/placement decisions with normal, borderline, and mildly retarded children. Face validity of the items in the domains included on the scale appears to be good. The type of derived scores are appropriate for classification/placement decisions. The ancillary measures provide safeguards against interpretation of invalid information. The primary type of information provided is related to social role performance outside of school from the perspective of the parent (or primary caretaker).

Although the ABIC is the best instrument published to date for assessment of adaptive behavior outside of school with normal or mildly handicapped children, a number of weaknesses should be recognized when interpreting scores. The age range is limited to 5-11 years. The norms are based entirely on California school age children. The accuracy of these norms in other settings and for other groups is questionable (Kazimour & Reschly, in press). An important domain of adaptive behavior for school age children, academic role performance, is not included on the scale, and is de-emphasized in Mercer's conception of adaptive behavior. Finally, practical considerations of time and resources may limit the implementation of this method of assessing adaptive behavior.

The Vineland Social Maturity Scale (VSMS) (Doll, 1953) is one of the oldest measures of social competence (adaptive behavior), and continues to be used quite widely (Coulter & Morrow, 1978). One of the reasons for the current use of the VSMS is that other scales are limited in age range or were not available until very recently.

The VSMS is a loosely structured interview which requires considerable skill on the part of the examiner. The VSMS attempts to measure social competence over the ages of birth to 30 years. As might be expected the items vary considerably in terms of sophistication and ease of administration. The domains of behavior covered by the VSMS are: Self-Help General, Self-Help Eating, Self-Help Dressing, Locomotion, Occupation, Communication, Self-Direction, and Socialization. The VSMS yields a composite score which can be transformed to a Social Quotient (SQ) which is a ratio of Social Age divided by chronological age and then multiplied by 100. The standard deviation of the SQ varies considerably from age to age which is generally the case with developmental scores such as ratio IQ, grade equivalents, etc. The norms for the VSMS are based on rather restricted samples of individuals assessed in 1935.

The VSMS is a venerable instrument which is in rather desperate need of revision and renorming, an activity that is currently underway which may substantially improve the scale. For older students it does provide some information that can be used to supplement clinical judgment of adaptive behavior. Direct use of SQ scores in classification/placement decisions is probably inappropriate for a variety of reasons (poor norms, limited sample of behavior, etc.).

The Children's Adaptive Behavior Scale (CABS) is a recently developed adaptive behavior scale which reflects some innovative approaches. The CABS (Richmond & Kicklighter, 1980) is administered directly to the child rather than to a third party respondent. The items on the CABS are organized around the rather typical domains of Language Development, Independent Functioning, Family Role Performance, Economic-Vocational Activity, and Socialization. In contrast to other adaptive behavior measures, the CABS appears to emphasize the cognitive competencies which are required

for various adaptive behaviors. For example, on the Independent Functioning Domain one of the items is "Where could you find a doctor?" In the Socialization Domain one of the items is "What should you say if someone gives you a piece of candy?" The norms for the CABS are based on rather restricted samples of slow learning and EMR students.

Relatively little is known about the CABS. It is likely that considerable research will be conducted with this procedure in the future. For the time being the CABS should be used cautiously if at all, pending research on its psychometric characteristics.

Research on Adaptive Behavior

Relatively little research has been published on the recently developed adaptive behavior scales. Three questions concerning adaptive behavior measures are particularly relevant to related services personnel. The limited evidence on these questions is reviewed in this section.

Relationship of Adaptive Behavior and Intelligence. A comprehensive review of the literature on social competence (the forerunner of adaptive behavior) and intelligence revealed a great deal of variability among studies (Leland, Shellhaas, Nihira, & Foster, 1967). The relationship between social competence and IQ varied depending on the measures used, the type of subject, and the variability within samples. However, in most studies, correlations between social competence and IQ were in the moderate range, about .4 to .6. These correlations, although substantial, indicate that social comptence and intelligence were quite different for a sizeable number of persons.

Relatively few studies of the correlations between IQ and recently developed measures of adaptive behavior have appeared in the literature. No studies were located for the ABS-Clinical or the ABS-PS. The significant differences on the ABS-PS between students in regular and EMR programs suggests that the ABS-PS is probably correlated at a statistically significant level with IQ. IQ scores were of course one of the bases for placing students in the EMR programs. However, these data do not provide information on the size of these correlations.

Correlations between the ABIC and WISC-R scores have been reported by a number of authors (Kazimour & Reschly, in press; Mercer, 1979; Oakland, 1980). These correlations have been in the low range varying from near zero to as high as .3 with a median of about .15. These correlations are considerably lower than those reported previously for social competence measures, and lower than the correlations reported by Mercer (1973, p. 187) for IQ and a forerunner of the ABIC used in the Riverside Studies. A number of reasons might account for these lower correlations. The most obvious factor is that the ABIC de-emphasizes the cognitive underpinnings of adaptive behavior and academic or school achievement types of behaviors are excluded. The evidence available to date suggests that the ABIC and measures of intelligence are largely independent.

In contrast to the results for the ABIC, fairly high correlations between adaptive behavior and intelligence were reported by the authors of the CABS (Kicklighter, Bailey, & Richmond, in press). For a sample of mildly retarded and slow learning children the correlations were in the range of .4 to .5. These correlations would probably be even higher if children from the full range of intelligence were studied. The reason for the higher correlations on the CABS in contrast to the ABIC is probably due to the greater emphasis on the cognitive aspects of adaptive behavior. It should be noted that the correlations for the CABS are more consistent

with the results of research conducted with the older measures of social competence.

The choice of adaptive behavior measures appears to be the major influence on the relationship of adaptive behavior and intelligence. Traditional measures such as the VSMS and the more recently developed CABS are correlated with intelligence at a moderate level. The correlation of the ABIC with intelligence is low enough that the relationship has no practical significance. The relationship of adaptive behavior to intelligence has significance for specification of the meaning of both constructs. Are the constructs independent? Are both a subset of a more general construct of general developmental level (Lambert, 1979)? Is one a subset of the other? If so, which is the more general construct? Theoretical formulations, research, and instrument development in the 1980's will undoubtedly address these questions.

Effects of Adaptive Behavior Measurement on Classification/Placement. Much of the impetus for the inclusion of adaptive behavior as part of a comprehensive assessment stemmed from concerns about overrepresentation of minorities in special classes for the mildly retarded. Depending on the adaptive measure used, research indicates that adaptive behavior assessment does indeed reduce overrepresentation of minorities in special classes for the mildly retarded (Reschly, in press; Talley, 1979). Fisher (1978) reported a high rate of "declassification" among all groups, not just minorities, as a result of the direct use of the ABIC in classification/placement decisions. The percentages of students declassified were 60, 70, and 85 for Anglo, Mexican-American, and Black students respectively. Apparently, use of the ABIC affects classification decisions with all groups, not just minorities. The question that remains is whether declassification is of benefit to the children involved.

The effects of other adaptive behavior scales on total percentages and group percentages of children classified as mildly retarded have not been reported in the literature. However, it is obvious that adaptive behavior and intelligence are far from being perfectly correlated. If very low scores on both dimensions are required for classification, then the prevalence of mild mental retardation among school age children will undoubtedly be well below the popular estimates of 2 to 2.5 percent. If the IQ cut off is at -2 standard deviations and adaptive behavior is based on out of school social role performance, the prevalence of all types of mental retardation among school age children will likely be closer to 1%, perhaps even lower. Assessment of adaptive behavior outside of school will have little if any effect on the prevalence of the more severe levels of mental retardation. The prevalence of moderate, severe, and profound levels of mental retardation, usually estimated at .3 to .5%, i.e., 3 to 5 per thousand, would be unaffected since persons obtaining IQs at these levels are nearly always found to be deficient in adaptive behavior as well (Grossman, 1977). The conception of and measurement procedures used to assess adaptive behavior have broad implications for the diagnostic construct of mild mental retardation.

Generalizability of Norms. Classification/placement decisions are typically made on the basis of degree of need, or the degree of deviation from typical patterns of behavior. Such decisions in the area of mildly handicapping conditions require the use of norm referenced measures. The representativeness and accuracy of norms for adaptive behavior measures is therefore an important consideration.

The situation with respect to the quality of the norms for existing adaptive behavior scales is not good. Both the ABS-PS and the ABIC use norms based exclusively on California children. The norms for the CABS and the VSMS are similarly restricted to persons from a specific geographic area along with other limitations.

The accuracy of the ABIC norms for children in other localities has been investigated on a limited basis. Buckley and Oakland (1977) studied the accuracy of California norms for two samples of Mexican-American children in Texas. The California mean scores were higher for both samples with a difference as large as 1/3 standard deviation for one of the samples. A difference of this magnitude might very well have implications for classification/placement decisions. Based on a study of three groups (Anglo, Black, and Mexican-Americans) in Texas, Gridley and Mastenbrook (1977) again concluded that California norms were inappropriate for Mexican-Americans, but acceptable for Anglos and Blacks. Kazimour and Reschly (in press) also found that the California means on the ABIC were higher in a study of four groups (Anglo, Black, Chicano, and Native American Papago) in Pima County Arizona. The size of the differences was rather small for all the groups except Native American Papagos whose ABIC composite mean was nearly 2/3 of a standard deviation below the California population average.

The available data suggest caution in use of the norms for adaptive behavior measures in other areas. The localities included thus far in studies have been restricted to the southwest. The generalizability of these findings to other areas is questionable. Even greater caution should be exercised in use of California norms with other sociocultural groups such as Native Americans, Southeast Asians, Orientals, etc.

Unresolved Issues in Adaptive Behavior

Although trite, the usual statement concerning the need for more research is clearly applicable to the area of adaptive behavior. A number of pressing issues are in need of resolution. We can only hope that the resolution that must take place during the 1980's will be guided by empirical evidence.

"Declassified" students. Use of existing adaptive behavior scales, particularly the ABIC, may lead to large numbers of students being "declassified," i.e., not being eligible for special education programs for the mildly retarded. Many others who would be eligible according to traditional criteria might not be placed in the future. Serious questions exist concerning whether these changes are beneficial to children.

To deny or ignore the educational problems experienced by the declassified children whould be naive and inhumane. Declassification in and of itself is a "nonsolution." Studies of the characteristics of children declassified through use of the ABIC have produced a fairly complex picture (Fisher, 1979; Scott, 1979). About half of the students were regarded as eligible for other special education classification. The other half were not eligible for existing special eduation services even though their intellectual and academic performance was well below average. Simply returning these students to regular classrooms, or avoiding special education classification with new referrals, does nothing about the aptitude and achievement problems. Special transitional programs have been funded for declassified students, but these are temporary and unrelated to the problems presented in new referrals (Meyers, MacMillan, & Yoshida, 1978; Yoshida, MacMillan, & Meyers, 1976). The concept of "permanent" transition programs stretches the imagination just a bit. The solution of this problem may be in refining the classification and in the type of special education services provided (see later section).

Student Role Performance and Adaptive Behavior. As we have seen the conception of adaptive behavior for school age children has been broadened in subsequent revisions of the AAMD Manual. We have moved from a point where academic achievement was the primary criterion to one current conception which completely ignores academic

achievement (i.e., the ABIC). Both positions are too narrow. Student role performance including academic achievement should be part of the conception of adaptive behavior for school age children (see later section).

Method of Measuring Adaptive Behavior. A traditional distinction among types of tests is the continuum of maximum performance vs typical performance instruments (Cronbach, 1970). Most traditional and current measures of adaptive behavior reflect a mixture of typical and maximum performance kinds of items. With typical performance measures the attempt is to determine how the individual customarily, habitually, or usually performs. The emphasis is not on "can" the individual perform the behavior, but rather on whether the individual "does" perform the behavior. The frequently used response choices on the ABIC of "usually," "occasionally," or "not at all" are good illustrations of this method of measurement.

A number of special problems exist with typical performance constructs and measures. The instruments are often subject to faking or other response sets. The ABIC attempts to control for response sets, but it is likely that the ABIC scores are not completely free of this kind of bias. Other adaptive measures do not control for response set biases. A second problem is the situation specific nature of many adaptive behaviors, particularly those which involve attitudes, social behaviors, or interpersonal competencies. Personality traits generally as well as many adaptive behaviors are likely to be exhibited by the individual in certain situations but not in others. To an unknown degree then an individual's adaptive behavior score is due to internal motivational and external situational contingencies. The degree to which the situation-specific factors in adaptive behavior measurement are a problem is determined largely by a third concern, the knowledge base of the respondent. The ideal situation would be a respondent who has opportunity and skills to thoroughly and accurately report on the child's behaviors in a wide variety of situations. Most respondents, even primary caretakers for children, do not have opportunities to observe children in all of the settings and roles included on adaptive behavior scales. The respondent's approach to those items where the knowledge base is incomplete may make a large difference. An acquiescent response set, independent of "faking good," may lead to spuriously elevated adaptive behavior scores. The acquiescent response set may operate in the following way. Consider an item in which the parent is asked to respond on whether the child acts as a helper in the classroom. Most parents' knowledge about this behavior is incomplete and second hand at best. One parent may acquiesce and say "Yes, he/she does that sometimes" while another may say that as far as they know the child never engages in that behavior. The problems are the limitations in respondent knowledge and the different approaches respondents may take to answering questions for which their knowledge is incomplete.

In addition to the problems discussed here, there are a number of other unresolved issues in the area of adaptive behavior. The interested reader is referred to Coulter and Morrow (1978) for a discussion of these issues.

Combining Intelligence and Adaptive Behavior Data

In addition to the other data from the multifactored assessment information on adaptive behavior and intelligence is particularly important in classification/placement decisions with the mildly retarded. How adaptive behavior is conceptualized and measured along with the available special education service options will have a significant influence on the classification/placement decisions that are made.

I suggest that the adaptive behavior dimension for school age children be conceptualized as two separate components. One component should involve performance

in the public school setting with primary emphasis on academic achievement in the classroom. The other component should be role performance in social systems outside of the public school such as the home, neighborhood, and community. Separation of the adaptive behavior dimension into two components is advisable because recently published data suggests that adaptive behavior in academic settings and social role performance outside of school are largely unrelated for many students (see previous discussion).

Inclusion of academic performance in the public school in our conception of adaptive behavior is consistent with the description of adaptive behavior for school age children in the AAMD Manual. Two of the nearly universal features of conceptions of adaptive behavior are age appropriate criteria and cultural context. Analysis of developmental task theory leads to recognition of the importance of academic performance during the ages of about 5 to 16 or 18 in our culture. Academic role performance is an important cultural expectation that is common to all major groups in our society. If adaptive behavior is "the way an individual performs those tasks expected of someone his(her) age in his(her) culture" then academic performance must be included in any comprehensive view of the construct of adaptive behavior for school age children.

Our conception of adaptive behavior should not be restricted to role performance in academic settings. Other social roles and other social systems are also important domains of development. Again the conception of adaptive behavior in the AAMD Manual and developmental task theory can be cited as foundations for this second component of adaptive behavior. During the school age years children perform a variety of social roles of increasing complexity in various social systems. To ignore the child's strengths and weaknesses in social systems outside of the school would also constitute a serious deficiency in our view of adaptive behavior.

Classification and placement decisions with the mildly retarded should be based on information from both components of adaptive behavior and the dimension of intelligence. Tables 8 and 9 provide a model for a two dimensional conception of adaptive behavior and a scheme for combining information on adaptive behavior in classification/placement decisions.

The different combinations of adaptive behavior and intelligence have implications for classification and placement decisions. Adaptive Behavior-School (AB-S) should be based on a complete educational evaluation including observation in the classroom, examination of samples of daily work, teacher interview, and the results of individually administered standardized achievement tests. Adaptive Behavior-Outside School (AB-OS) should be based on information from formal inventories such as the ABIC, where appropriate, or informal data collection procedures.

Of particular interest are the children who exhibit the pattern of very low intelligence, very low AB-S, and normal AB-OS. A major current dilemma is whether these children should be classified and placed in special education programs. Such children are almost by definition "Six Hour Retarded Children." If they are classified and placed in special education programs we will almost inevitably overrepresent minority children. In my view these children should be served in special education programs in most instances because they do in fact have extreme educational needs that are typically beyond the scope of regular classroom instruction. The solution of "delabeling" these children does not address these needs. However, the segregated special class for the mildly retarded which has often been the placement used, because in many cases it was the only alternative, is an equally inappropriate solution.

Table 8

Conception of Adaptive Behavior for School Age Children

ADAPTIVE BEHAVIOR: SCHOOL BASED

Rationale: 1) Mastery of literacy skills is a key developmental task for persons between the ages of 5 and 17

2) The expectation for and emphasis on educational competencies is common to most if not all major sociocultural groups.

Assessment: 1) Collection and consideration of a broad variety of information including teacher interview, review of cumulative records, examination of samples of classroom work, classroom observation, results of group standardized achievement tests, results of individual achievement tests, diagnostic achievement tests, and other informal achievement measures.

ADAPTIVE BEHAVIOR: OUTSIDE OF SCHOOL

Rationale: 1) Mastery of a variety of non-academic competencies also is expected, and a key developmental task between the ages of 5 and 17.

2) The expectations for and opportunities to develop non-academic competencies may vary among sociocultural groups.

Assessment: 1) Collection of information on social role performance outside of school in areas such as: peer relations, family relationships, degree of independence, responsibilities assumed, economic/vocational activities, etc.

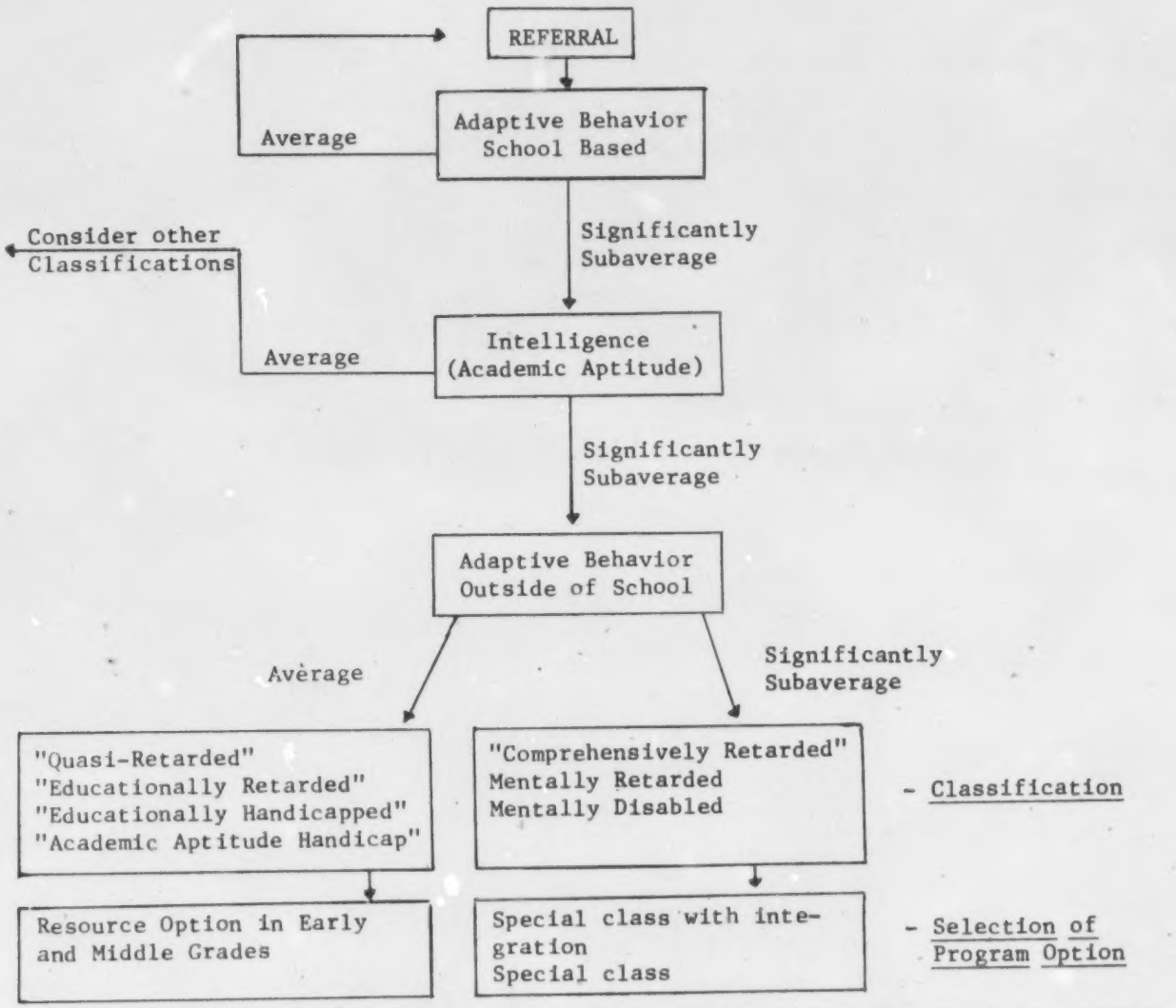
2) Method of collecting data may include formal measures, interviews with parents, interview with student, etc.

The solution to this dilemma depends on two developments. First, we need a more refined classification system which would differentiate between what Mercer (1973) called the Quasi and Comprehensively Retarded. According to Mercer's scheme the Comprehensively Retarded are persons who fail both components of the adaptive behavior dimension and the intelligence dimension. The Quasi-Retarded exhibit the same pattern except for normal social role performance outside of school. The overrepresentation of minorities in special education classes for the educable mentally retarded in the Riverside, California schools (Mercer, 1973), and in other locations as well, is largely attributable to placement of Quasi-Retarded children who come from minority backgrounds. Should these children be labeled as mentally retarded? Opinions on this issue differ sharply (Goodman, 1979; Mercer, 1979).

A refinement in the classification system would be beneficial in resolving this dilemma. The terms Comprehensive and Quasi are probably objectionable to many as is the term mental retardation. Use of the terms Educational Retardation, Educationally Handicapped, or some other term which is as behaviorally descriptive as possible of the Quasi-Retarded pattern would be preferable. Greater refinement in the classification system is useful only if there are implications for placement decisions and educational programming. The change suggested may have such implications.

Table 9

A Tentative Scheme for Use of Adaptive Behavior Information in Classification and Selection of Program Option



The "Quasi-Retarded" do need special services. However, if special education services are to be provided, the objectives should be oriented toward specific academic needs rather than broad social competencies. In most instances the resource program involving remedial and compensatory tutorial services is a more appropriate option than the special class. Special class programs for the mildly retarded have traditionally placed considerable emphasis on broadly defined social competencies and "functional" academic skills (Kolstoe, 1976). This emphasis is clearly appropriate for the comprehensively retarded, but is probably misdirected for most of the Quasi-Retarded. With few exceptions the Quasi-Retarded, if placed in special education, should be placed in resource programs.

Use of the resource option for the Quasi-Retarded would alleviate many of the concerns expressed by Federal District Courts in the placement litigation. The

amount of time spent outside of the educational mainstream is minimized by the resource option thus reducing the very proper concern about racial segregation. Placement in the resource option regardless of classification used may have the additional advantage of being less stigmatizing. Analysis of outcome data must, of course, be the ultimate criteria against which this or any other classification/placement system must be validated.

Refined classification decisions along with selection of service option, resource vs special class, appear to be promising applications of adaptive behavior assessment. Other applications of adaptive behavior data with the mildly handicapped are also promising. General strengths and weaknesses across different domains of behavior may be the initial source of information for developing interventions designed to improve social skills, assertiveness, etc. The information from currently available instruments such as the ABIC is not sufficiently precise for direct translation to intervention objectives. Data from the ABIC, AAMD-PS, or Vineland can alert us to general needs which can then be translated to specific objectives through additional observation and/or interview.

SOCIOCULTURAL BACKGROUND

The PL 94-142 Rules and Regulations list social or cultural background as one of the areas that "shall" be considered in placement decisions. The apparent purpose of the regulation is to ensure that socioeconomic and cultural factors are considered in interpreting information from other sources. Consideration of such factors was suggested in the placement litigation (e.g., Guadalupe case) of the early 1970's where in several instances bilingual children were allegedly misplaced in special class programs for the mildly retarded. Nearly everyone would agree that social, economic, and cultural background factors should be assessed and considered in classification/placement decisions. In extreme situations, e.g., Southeast Asian students who have recently emigrated to the United States, most would agree that conventional measures should be interpreted in light of sociocultural factors, and that special education classification/placement decisions should be delayed until the child has a chance to learn the language of the school, become familiar with American culture, etc. Such children may well need special services, but conventional special education classifications are inappropriate. The difficult issues in this area are the consideration and use of such data with native born Americans who are to varying degrees different from the majority population on social and cultural variables. The major questions are how to assess the sociocultural variables and how this information should be used in classification/placement AND educational programming decisions. As is the case with several of the Federal Rules and Regulations, there is no elaboration or guidelines for the measurement and use of sociocultural information. Furthermore, in contrast to adaptive behavior, there are even fewer resources in psychology or education in terms of theory, research, and instruments that can be applied to the area of sociocultural factors.

The Concept of Eth-class

The concept of sociocultural background includes the overlapping factors of social class and race or ethnicity. Mercer (1979) used the term eth-class from sociology to refer to these combined effects. The concept of eth-class, or the more commonly used term of sociocultural background, is needed to accurately describe the variations among groups on measures of intelligence or achievement.

Socioeconomic Status (SES). Social class differences exist within, and unfortunately, between major ethnic-racial groups in the United States. These differences

are associated with a variety of conditions related to economic resources, educational level, attitudes and values, religious and political preferences, etc. In short, social class differences influence the individual's lifestyle and opportunities. Most pertinent to this chapter are the discontinuities between "the middle class teacher and the every-class child" (McCandless, 1967). The interested reader is referred to the excellent discussion by McCandless of the practical significance of social class in terms of child development and education.

Measures of social class vary from relatively simple occupational scales to four or five factor indices based on occupation, educational level, source of income, housing type, and area of residence. In most published research the measure of social status is typically based on occupation and educational level of the child's parent(s). These two sources of information are relatively easy to obtain and are closely related to the results of the more thorough measures of SES.

That SES is related to measured intelligence has been known at least since the early years of this century. The relationship is far from perfect. The correlations between SES and intelligence or achievement are typically in the range of .3 to .4; the range of performance within SES levels is fairly large; and considerable overlap of distributions is typical. The relationship of SES to average levels of intelligence appears to be more impressive. For example, Kaufman and Doppelt (1976) reported mean differences of 9 to 17 points for both blacks and whites between the highest and lowest SES groups in the WISC-R standardization sample. Multiple correlations in the .30's and .40's have been reported between WISC-R IQs and race, sex, and SES (Reynolds & Gutkin, 1979) with SES being the best predictor of intellectual level.

SOMPA Sociocultural Scales. The Sociocultural Scales (SC) in SOMPA are more sophisticated than traditional measures of socioeconomic status. Some information on cultural background is also included. The SC are based on 22 questions (24 items) which are organized into nine factors and four sociocultural modalities. The modalities, factors, and type of information gathered through the SC are presented in Table 10. The SC are administered to the primary caretaker of the child in an interview that also includes the SOMPA ABIC and SOMPA Health History Inventory.

Table 10

SOMPA Sociocultural Scales

Modality	Factor(s)	Type of Information
Family Size	Family Size	Number of siblings and number of persons in the household
Family Structure	Parent-Child Relationship Marital Status	Relationship of child to parents, gender of head of household, and marital status
Socioeconomic Status	Occupation Source of Income	Kind of work of head of household and source of income for family
Urban Acculturation	Sense of Efficacy Community Participation Anglicization Urbanization	Agree-disagree statements on what determines success, the value of planning, and delay of gratification. Questions on amount and kind of community participation. Educational level of parent, location and size of place where parents were reared, and a rating of the respondents mastery of English.

The items on the SC are based on published research concerning factors related to measured intelligence. The correlations of the factors and modalities differ in size within and between ethnic groups. The multiple correlations between the SC and WISC-R Full Scale IQs vary from .37 to .42 depending on group (Anglo, Black, or Hispanic) (Mercer, 1979, Table 44). The correlation of the four modalities with the Full Scale IQ score varies for different groups. Socioeconomic status has the highest correlation for Anglos (.39) while Urban Acculturation is highest for Blacks and Hispanics (.30 and .37 respectively). The Family Structure Modality has relatively low correlations with the WISC-R for Anglos and Blacks (.13 and .15 respectively) and is not correlated with any of the WISC-R scores for Hispanics.

Mercer's argument for pluralistic norms (see below) was bolstered by the data on the relationship of the SC to the WISC-R. She suggested three criteria indicating the need for pluralistic norms: 1) Significant differences among groups on measures of intelligence; 2) Significant differences among groups in sociocultural measures; and 3) Sociocultural measures account for a significant amount of the variation in measured intelligence within and between groups. These criteria were met in her studies of California school age children representing three groups (Anglo, Black, and Hispanic). The subsequent development of pluralistic norms has become the most controversial aspect of SOMPA.

SOMPA Estimated Learning Potential. The SOMPA Estimated Learning Potential (ELP) procedure is the formal method developed by Mercer to eliminate the biases in IQ tests. A multiple regression equation using the Sociocultural Scales (SC) as predictors and the WISC-R IQ scales as criteria were developed for the three groups in the California standardization sample. Separate regression equations are used for each group. Although seemingly complex, the entire procedure simply involves changing the WISC-R mean and standard deviation to 100 and 15 respectively for all groups, and then computing individual scores through differential weighting of the four SC modalities. The amount of change for any individual within each of the groups depends on his/her sociocultural characteristics. The net effect is to remove group differences through an algebraic transformation. The question now, as posed by one of the commentators on SOMPA, is "The Algebra Works - But What Does It Mean?" (Brown, 1979).

Before reviewing research and commentaries on the ELP it perhaps is important to recognize that Mercer's ELP procedure is not the first time that someone suggested changing IQs depending on the child's background. Platt and Bardon (1967) quoted Havighurst (1951) as recommending "A good rule to follow is to add ten points to the IQ of all children who come from underprivileged homes or homes where English is not spoken as the first language." The reasons for this sort of adjustment of scores have never been altogether clear. They seem to be related to myths that IQs should represent innate ability. That myth continues to haunt our efforts and has been a part of the SOMPA debate. The ELP does represent a much more systematic and logical method for adjusting scores.

Meaning of ELP scores. One of the obvious questions about use of the SOMPA ELP scores is whether the California norms can be generalized to other areas and to other sociocultural groups, e.g., Native Americans. The SC data and ELP regression equations provided by SOMPA are based entirely on the SOMPA Standardization sample of California school age children. The California normative data will not be accurate for other areas or groups if there are significant differences in mean WISC-R scores, in sociocultural characteristics, or in the relationship of the SC to WISC-R scores. Mercer (1979) suggests that local norms be developed through studies of representative samples of normal children. The minimum sample size recommended is 25 males and 25 females at each of 7 age levels, or a total sample of

350 students (Mercer, 1979, p. 144). Data to be collected includes the WISC-R and the SC. In addition, fairly complex statistical analysis are required. Due to limitations in resources it is highly unlikely that studies meeting these criteria will be conducted in very many localities or for very many groups.

Studies on the generalizability of ELP scores have been conducted with data sets from Pima County, Arizona and Austin, Texas (Reschly, 1980; Oakland, 1980). The limited data available now suggests caution in using the California ELP norms in other areas or with other groups. The studies that have been conducted involved samples from the Southwest. These populations are likely to be more like California samples than say samples of Hispanic children from New York City or samples of black children from South Carolina.

A second major concern has to do with the possible effects of the use of ELP scores in classification/placement decisions. A number of authors have expressed strong reservations about using ELP scores in classification/placement decisions (e.g., Clarizio, 1979). However, this is the primary purpose of the ELP concept, and these scores will undoubtedly be used by some in classification/placement decisions.

For all children, the ELP score is either the same as or higher than the conventional WISC-R IQs. The magnitude of these differences sometimes is quite large. Direct use of the ELP score will therefore have implications for children classified as mildly retarded. Fisher (1978) reported that 40 to 75% of the children currently classified as EMR would be "declassified" if the ELP score was used rather than the conventional score. The greatest effect was on minority students with little change noted for Anglos. The obvious question that remains is whether the possible "declassification" due to use of ELP will be beneficial to students.

A third issue has to do with the validity of the ELP scores. Mercer (1979) argues strongly that the validity of ELP must be assessed in the context of the values and purposes of her Pluralistic Assessment Model. From that perspective, the ELP is valid in that the differences between sociocultural groups are eliminated, and variations within groups are preserved. The broader perspective adopted by critics is that ELP in order to be useful must relate to other criteria. The question is, which criterion? The relationship of ELP to achievement test scores or teacher ratings of achivement is not as strong as the relationship of the conventional scores (Oakland, 1980; Reschly, 1978). Mercer rejects these data as irrelevant to the construct of ELP. In her view the key is not the relationship of ELP to past achievement, but rather the degree to which ELP would predict acquisition of new material, or learning rate. The technology available to assess learning potential or learning rate is not well developed, or easily applied. This kind of study, i.e., relating ELP to learning rate, is needed in order to establish the predictive validity of ELP as well as clarify a number of fascinating theoretical issues.

One of the most intense debates concerning SOMPA is over the issue of separating ignorance from stupidity (Goodman, 1979; Mercer, 1979). One of the purposes of the ELP is to determine whether the child is "stupid" or merely "ignorant." This clearly borders on the notion of attempting to separate innate potential from current level of functioning, or true mental retardation from pseudo-mental retardation. Children with low conventional scores and high ELPs are presumed to be ignorant while those who score low on both might be regarded as stupid (or "truly" retarded). This argument is probably beyond resolution through empirical study, although data on the ELP learning rate relationship would be interesting in this context. The broader issues in the ignorance-stupidity debate are the meaning of mild mental retardation and the meaning of IQ test scores.

Although much of the discussion of ELP in this chapter has been skeptical, the ELP concept may be highly useful in one important area, clarifying the meaning of IQ test results. In SOMPA Mercer renames the conventional WISC-R scores as School Functioning Level (SFL). Although I would prefer the term academic aptitude, the result is virtually the same. Renaming what the tests measure may reduce misconceptions about IQ test results.

PRIMARY LANGUAGE

The assessment of primary language competence is a logical, common sense procedure as well as a requirement in the recent legislation. Non-English speaking children have apparently been placed in programs for the mildly retarded on the basis of tests administered in English (see Diana or Guadalupe cases). These classification and programming decisions were inappropriate, although an even larger problem in those situations was the apparent absence of alternative programs for non-English speaking youth.

Assessment of primary language competence is more difficult than it might appear. Many instruments have been developed recently (see Oakland, 1977), but little systematic work has been conducted on their reliability and validity. Nevertheless, systematic effort to assess primary language competence is needed. The decision about primary language competence must be based on data. The presence of a Latino surname, for example, is certainly not sufficient to conclude that the child or family uses Spanish as the dominant language. The author is acquainted with cases of Latino surnamed families where Spanish is not spoken, and has not been used in the family for several generations. Conversely, the author encountered a case in 1967 in eastern Iowa where the child had an Anglo surname, but was monolingual Spanish speaking.

The information on primary language is important in collecting and interpreting other assessment data, and in decisions about appropriate interventions. If the child is monolingual, non-English speaking, perhaps the wisest course of action is to simply avoid the use of norm referenced standardized tests of achievement and ability. The 94-142 regulations suggest use of an interpretor. Due to the many problems which arise when attempts are made to translate tests into other languages, e.g., items do not have the same meaning and difficulties of items change, the results of translated tests are of questionable value. If inferences must be made about ability, use of nonverbal or performance tests is probably the best course of action. Educational programs for monolingual non-English speaking students should be provided in the students' native language if at all feasible. If only a few monolingual children attend schools in a particular district, then other alternatives should be pursued (see Oakland, 1977).

Bilingual children may exhibit widely varying competencies in English and another language. The range will extend from limited to high degrees of competence in either or both languages. The language dominance measure that is used to determine primary language should be supplemented by other measures which yield information on competence in both languages. Subsequent assessment activities should be conducted witin the dominant language of the child. An important principle to remember is the assumption of maximum performance. Any inference about ability or academic aptitude made in subsequent assessment activities should include consideration of the effects of differences in language. Bilingual youth may, though certainly not always, obtain lower scores on verbal measures administered in English due to limited exposure to English. Special education services may not be the appropriate intervention for bilingual children who, on the basis of other data, meet the state guidelines for special education classification. Bilingual/bicultural

programs may be more appropriate, and children's rights to such services have been established through recent litigation.

MILD MENTAL RETARDATION: "A CONTINUING DILEMMA"

Much of the professional debate, litigation, and legislation over bias in assessment involves implicit and contradictory assumptions about the nature of mental retardation. The meaning of mild mental retardation, called "a continuing dilemma" by Zigler (1967), has been a particular problem. Consensus regarding the meaning of this diagnostic construct would greatly assist efforts to resolve the issues discussed in this paper.

Definitions and Classification Criteria

Terminology and classification criteria in mental retardation have evolved throughout the present century. There are two major sources of terminology and classification criteria which are crucial for diagnostic personnel in the public schools. The American Association on Mental Deficiency (AAMD), the major professional organization in the field, publishes a terminology and classification manual. The AAMD Manual is revised periodically with the most recent revisions published in 1977, 1973, and 1961 (Grossman, 1973; 1977 and Heber, 1961). The 1973 and 1977 revisions are virtually identical. The AAMD Manual on Terminology and Classification has a significant influence on other definitions and classification criteria in mental retardation. The influence of 1961 and 1973 versions are to varying degrees reflected in state education codes, the second major source of guidelines for terminology and classification in mental retardation. State education codes usually provide a definition and classification criteria for mental retardation which are to be applied by public school diagnostic personnel. Although the AAMD system is important, it should be noted that decisions in the public schools are to be based on the state definition and criteria for mental retardation. The terminology, classification criteria, etc., for mental retardation vary considerably among states (Patrick & Reschly, 1980). Knowledge of your current state code, usually published in the State Special Education Rules and Regulations, is a necessity for diagnostic personnel.

MacMillan (1977) and Robinson and Robinson (1976) provide thorough analyses of the AAMD classification system in mental retardation. Some of the most important characteristics of the AAMD system are the following.

1. Bi-Dimensional. The individual must exhibit deficits in both intelligence and adaptive behavior in order for the classification of mental retardation to be appropriate.

2. Developmental. The deficits in intelligence and adaptive behavior must appear during the developmental period which is defined as the ages of birth to 18.

3. Current Status. "Mental retardation is descriptive of current behavior and does not necessarily imply prognosis" (Grossman, 1977, p. 11).

4. Etiology. Etiology of mental retardation is not specified in the definition. Etiology may be due to psychosocial, psychogenic, or biological influences.

5. Continuum. All types and levels of mental retardation are implicitly organized on the same continuum ranging from mild to profound.

6. Levels. The level of severity of mental retardation is specified by standard deviation (s.d.) cut off points for IQ test scores.

Mild (Educable)	- 2 s.d.	IQ Range of about 55-69
Moderate (Trainable)	- 3 s.d.	IQ Range of about 40-54
Severe	- 4 s.d.	IQ Range of about 25-39
Profound	- 5 s.d.	IQ Range of about 24 and below

7. Adaptive Behavior. The criteria for adaptive behavior depends on the age of the person. (See earlier discussion.)

The AAMD classification scheme depicts mental retardation as a current status with no implications for etiology or prognosis. Further, the notion of mental retardation as situational (vs comprehensive) incompetence is at least implied in the different criteria for adaptive behavior depending on the age of the individual. Clearly, the AAMD classification scheme does not require that mental retardation be a permanent status, or be due to biological anomaly.

Much of the litigation as well as other discussions of bias in assessment reflect the implicit misconception that mental retardation requires permanent incompetence, comprehensive incompetence, and biological anomaly. Mercer's report on the Riverside studies (Mercer, 1973), the Larry P. Opinion, and the concern for "six hour" retarded children are examples of these misconceptions. "Six hour" retarded children were described in a 1970 President's Committee on Mental Retardation report as retarded only in the public school context, thus the adjective "six hour." In other social settings they were described as coping in ways that "...may be exceptionally adaptive to the situation and community in which they live." Should these students, who are failing in the classroom, have low intelligence and achievement scores, etc., be classified as mentally retarded?

The answer to this question obviously varies according to what the diagnostic construct of mental retardation means. The AAMD Manual and most state education codes would allow classification in the mild or educable level of mental retardation. Whether such children are "truly" retarded, i.e., permanent and comprehensive impairment due to biological anomaly, is largely irrelevant in these classification systems. They may be classified as mildly retarded on the basis of serious problems in the classroom, low intelligence and achievement scores, etc. However, from Mercer's perspective as well as that of the courts and the Federal Office for Civil Rights, these children are not "truly" retarded.

The current debate over true vs pseudo or quasi retardation is reminiscent of the earlier discussion of pseudofeeblemindness (Benton, 1956). The 1961 and subsequent revisions of the AAMD Manual represented attempts to avoid the issues of precise etiology (which usually is unknown) and prognosis (which often is unclear). However, there is an implicit problem in the AAMD system which contributes to the confusion over the meaning of mental retardation. The mild (or educable) level of mental retardation is markedly different from the moderate, severe, and profound levels on a number of dimensions. Among these dimensions are:

1. Etiology. The presumed etiology for most cases of mild mental retardation is the AAMD category of psychosocial. The vast majority of the mildly retarded do not exhibit any evidence of biological anomaly. In contrast, the more severely retarded almost always have biological anomaly, although the precise etiological mechanism often is unknown.

2. Age. Mild mental retardation is rarely diagnosed prior to 5 or 6, and the highest prevalence is usually found in late childhood or early adolescence. In contrast, most cases of moderate, severe, or profound mental retardation are diagnosed during the preschool years, usually prior to age 2.

3. Comprehensive. The behavioral deficits of children who are classified as mildly retarded are usually restricted to the public school setting. Performance in other settings is usually regarded as normal by significant others such as parents, siblings, and other adults. The moderately, severely, and profoundly retarded are typically regarded as retarded in nearly all settings.

4. Socioeconomic Status (SES). There is a strong association between socioeconomic status and mild mental retardation. Children who are diagnosed as mildly retarded are much more likely to come from low SES environments. The relationship between SES and other levels of mental retardation is very weak, if it exists at all.

5. Ethnic/Racial Status. The prevalence of mild mental retardation is higher among specific ethnic/racial groups if the group is also of lower SES. The more severe levels of mental retardation are not found with any greater frequency among specific ethnic/racial groups.

6. Permanence. Most of the persons diagnosed as mildly retarded become independent functioning, self-supporting adults (Baller, Charles, & Miller, 1967). The diagnosis of mild mental retardation is therefore not permanent for these individuals since their adaptive behavior during adulthood is within normal limits. Again in contrast, nearly all of the more severely retarded are unable to function with complete independence or become entirely self-supporting at any time during their life span. At the more severe levels (moderate, severe, and profound) mental retardation is, almost without exception, a permanent condition.

The implicit problem in the AAMD system is that all levels of mental retardation are placed on the same continuum despite the differences cited above in the mild vs the more severe levels. The adjectives mild, moderate, etc., have not been effective in communicating these differences. Another way of analyzing this problem is to consider the connotative and denotative meanings of mental retardation. The denotative, or precise scientific, meaning of mental retardation is restricted to current status with no assumptions about etiology or prognosis. However, the connotative (everyday, lay public) meaning of mental retardation is that of comprehensive incompetence and permanent disability due to biological anomaly. Changes in terminology and conceptions of diagnostic constructs are indicated when the associated connotative and denotative meanings are widely divergent. This appears to be the case with mental retardation.

Revision in the mental retardation classification system in the form of clear separation of mild from other levels of mental retardation would aid in solving this problem. Terms such as educationally retarded or academically handicapped

would be more appropriate descriptors of the kinds of problems displayed by students classified now as mildly retarded. Other terms might be better than the examples used here. The point is to reduce the miscommunication and misconception about mental retardation. Revisions in the classification system would assist in resolving some of the issues discussed in this paper. Such revisions, however, are not panaceas.

NONBIASED ASSESSMENT: SOME TENTATIVE CONCLUSIONS

Nonbiased assessment is obviously an extremely complex issue. Concerns with the meaning and usefulness of IQ test results have dominated much of the discussion of nonbiased assessment. The issues surrounding the meaning of IQ (academic aptitude) have been debated for at least sixty years, and are not likely to be resolved in the near future. However, many other issues such as the meaning and etiology of mild mental retardation, the rights of parents and students, the effectiveness of special education interventions, and the definition of bias in tests are clearly involved with our efforts to reduce bias in assessment. These issues have been discussed in this paper, though certainly not resolved.

There are two possible reactions among a range of possible reactions to the pressures for nonbiased assessment which could be damaging to children. One possible reaction is to conclude that the issue is so complex and ill defined that there is nothing we can do, hence, we should stubbornly defend and simply continue our current practices. This reaction will be maladaptive. There are important changes that we can make which will enhance the fairness and usefulness of assessment for all children. In the interests of children, we need to make these changes. A second maladaptive reaction is to reject most if not all of our current instruments and practices. For example, some have rejected the use of IQ tests with culturally different children. Others have severly limited the numbers of culturally different children in special education programs simply on the basis of their proportions in the population. Such reactions are not in the best interests of children.

Positive reactions to the concerns about nonbiased assessment must first be based on a recognition of the ambiguity of the current situation. There are no and probably never will be any easy solutions.

Recognition of the underlying assumptions in the special education placement litigation provides an orientation to the most important issues in nonbiased assessment. One can only wonder if these cases would have appeared IF the interventions were effective; IF due process safeguards had been observed; IF the interventions had been consistent with the principle of least restrictive alternative, i.e., had not been provided in segregated, self-contained special classes; IF the assessment had been multifactored and programs based on specific educational need; and so on. The fact is that assessment and programs did NOT meet these criteria in at least some, and perhaps, many instances. The litigation and legislation are attempts to correct these abuses. From the perspective of related services personnel and special educators, the current demands for nonbiased assessment along with the other requirements from the courts and legislation, are the best things that have happened for our professions (and for children).

Three general themes should form the basis for efforts to achieve nonbiased assessment. First, and most important, we must continue and expand our efforts to insure that assessment procedures result in positive benefits for individuals. This goal is certainly not new. The underlying assumption of positive benefit to individuals has always been the goal in all types of assessment. Realization of this goal requires more concern about the relationship of our assessment activities

to interventions, and more concern about the effectiveness of these interventions.

A second theme is the need to implement the idea of multifactored assessment. Again, this is not a new idea. However, the degree to which comprehensive assessment was conducted, documented, and used in planning interventions has varied considerably. The proper role of IQ tests in the multifactored assessment must be recognized. Areas often ignored in the past, e.g., adaptive behavior outside of school, primary language competence, and sociocultural background, should be a part of the assessment process. These newer areas of assessment, along with the conventional areas, are important to better understanding of children. Fuller understanding can lead to better, more refined classification decisions and more effective interventions.

Finally, our understanding of nonbiased assessment and our ability to implement these procedures will be enhanced if we view nonbiased assessment as a process rather than a set of instruments. The process is oriented toward insuring fairness and effectiveness of assessment and interventions for all children. The process is appropriate in all settings regardless of the ethnic or racial composition of the student population. The nonbiased assessment process is perhaps best illustrated by the series of questions developed by the Northeast Regional Resource Center. A copy of this document is included in an appendix. A second guideline which follows was developed by a committee appointed by the Iowa Department of Public Instruction. Both documents are attempts to identify key features of a nonbiased (and effective) assessment process.

APPENDIX A

PROTECTION IN EVALUATION PROCEDURES PROVISION OF PL 94-142

PROTECTION IN EVALUATION PROCEDURES

§ 121a.530 General.

(a) Each State educational agency shall insure that each public agency establishes and implements procedures which meet the requirements of §§ 121a.530–121a.534.

(b) Testing and evaluation materials and procedures used for the purposes of evaluation and placement of handicapped children must be selected and administered so as not to be racially or culturally discriminatory.

(20 U.S.C. 1412(5)(C).)

§ 121a.531 Preplacement evaluation.

Before any action is taken with respect to the initial placement of a handicapped child in a special education program, a full and individual evaluation of the child's educational needs must be conducted in accordance with the requirements of § 121a.532.

(20 U.S.C. 1412(5)(C).)

§ 121a.532 Evaluation procedures.

State and local educational agencies shall insure, at a minimum, that:

(a) Tests and other evaluation materials:

(1) Are provided and administered in the child's native language or other mode of communication, unless it is clearly not feasible to do so;

(2) Have been validated for the specific purpose for which they are used; and

(3) Are administered by trained personnel in conformance with the instructions provided by their producer;

(b) Tests and other evaluation materials include those tailored to assess specific areas of educational need and not merely those which are designed to provide a single general intelligence quotient;

(c) Tests are selected and administered so as best to ensure that when a test is administered to a child with impaired sensory, manual, or speaking skills, the test results accurately reflect the child's aptitude or achievement level or whatever other factors the test purports to measure, rather than reflecting the child's impaired sensory, manual, or speaking skills (except where those skills are the factors which the test purports to measure);

(d) No single procedure is used as the sole criterion for determining an appropriate educational program for a child; and

(e) The evaluation is made by a multidisciplinary team or group of persons, including at least one teacher or other specialist with knowledge in the area of suspected disability.

(f) The child is assessed in all areas related to the suspected disability, including, where appropriate, health, vision, hearing, social and emotional status, general intelligence, academic performance, communicative status, and motor abilities.

(20 U.S.C. 1412(5)(C).)

Comment. Children who have a speech impairment as their primary handicap may not need a complete battery of assessments (e.g., psychological, physical, or adaptive behavior). However, a qualified speech-language pathologist would (1) evaluate each speech impaired child using procedures that are appropriate for the diagnosis and appraisal of speech and language disorders, and (2) where necessary, make referrals for additional assessments needed to make an appropriate placement decision.

§ 121a.533 Placement procedures.

(a) In interpreting evaluation data and in making placement decisions, each public agency shall:

(1) Draw upon information from a variety of sources, including aptitude and achievement tests, teacher recommendations, physical condition, social or cultural background, and adaptive behavior;

(2) Insure that information obtained from all of these sources is documented and carefully considered;

(3) Insure that the placement decision is made by a group of persons, including persons knowledgeable about the child, the meaning of the evaluation data, and the placement options; and

(4) Insure that the placement decision is made in conformity with the least restrictive environment rules in §§ 121a.550–121a.554.

(b) If a determination is made that a child is handicapped and needs special education and related services, an individualized education program must be developed for the child in accordance with §§ 121a.340–121a.349 of Subpart C.

(20 U.S.C. 1412(5)(C); 1414(a)(5).)

Comment. Paragraph (a)(1) includes a list of examples of sources that may be used by a public agency in making placement decisions. The agency would not have to use all the sources in every instance. The point of the requirement is to insure that more than one source is used in interpreting evaluation data and in making placement decisions. For example, while all of the named sources would have to be used for a child whose suspected disability is mental retardation, they would not be necessary for certain other handicapped children, such as a child who has a severe articulation disorder as his primary handicap. For such a child, the speech-language pathologist, in complying with the multisource requirement, might use (1) a standardized test of articulation, and (2) observation of the child's articulation behavior in conversational speech.

§ 121a.534 Reevaluation.

Each State and local educational agency shall insure:

(a) That each handicapped child's individualized education program is reviewed in accordance with §§ 121a.340–121a.349 of Subpart C, and

(b) That an evaluation of the child, based on procedures which meet the requirements under § 121a.532, is conducted every three years or more frequently if conditions warrant or if the child's parent or teacher requests an evaluation.

(20 U.S.C. 1412(5)(C).)

From p. 42496 & 42497 of Federal Register, August 23, 1977. Education of Handicapped Children, Regulations Implementing Education for All Handicapped Children Act of 1975. (Public Law 94-142).

APPENDIX B

OUTLINE OF NONBIASED ASSESSMENT PROCEDURES DEVELOPED BY THE NORTHEAST REGIONAL RESOURCE CENTER

REFERRAL

1. Are the parents/guardians aware that a referral has been made for their child, and by whom?

2. Is this child's presenting problem clearly and precisely stated on the referral?
   a. Does the referral include descriptive examples of behavior rather than opinions of the referring agent?
   b. Is there supportive documentation of the problem?

3. Is the referral legitimate?
   a. Does the referring agent have a history of over referral of children from certain cultural groups?
   b. Could irrelevant personal characteristics (e.g., sex or attractiveness) of the child have influenced the decision to refer him?
   c. Could the referring agent have misinterpreted this child's actions or expression due to his lack of understanding of cultural differences between himself and the child?

4. Can the assessment team provide the referring agent with interim recommendations that may eliminate the need for a comprehensive evaluation?
   a. Is it possible that the curriculum being used assumes that this child has developed readiness skills at home that in reality he hasn't had the opportunity to develop? If so, can the team assist the teacher in planning a program to give this child the opportunity to develop readiness skills?
   b. Can the team provide information on the child's cultural background for the referring agent so that there are fewer misunderstandings between the referring agent and this child and perhaps other children of similar cultural background?

5. Have I informed this child's parents/guardians in their primary language of the referral?
   a. Have I explained the reason(s) for the referral?
   b. Have I discussed with the parents what next step activities may be involved?
      e.g., - professional evaluations
      - use of collected data
      - design of an individualized educational plan, if necessary
   c. Have I discussed due process procedures with the parents?
   d. Do I have documented parental permission for the evaluation?
   e. Have I asked the parents to actively participate in all phases of the assessment process?
   f. Have I informed the parents of their right to examine all relevant records in regard to the identification, evaluation and educational plan of their child?

MEETING THE CHILD

1. What special conditions about this child do I need to consider?
   a. What is the child's primary home language?
   b. Do I know about the child's home environmental factors?
      e.g., - familial relationships/placement
      - social and cultural customs

c. Do I understand this child's cultural and language so that I can evoke a level of performance which accurately indicates the child's underlying competencies?
d. Is this child impeded by a handicap other than the referral problem that may result in his not understanding what I am talking about?

2. What special conditions about myself do I need to consider?
a. How do I feel about this child?
b. Are my values different from this child's?
c. Will my attitude unfairly affect this child's performance?
d. Can I evaluate this child fairly and without prejudice?
e. If not, would I refer him to another assessor if one is available?

3. Have I examined closely all the available existing information and sought additional information concerning this child?
a. Has the child's academic performance been consistent from year to year?
b. Is there evidence in this child's record that his performance was negatively or positively affected by his classroom placement or teacher?
c. Are his past test scores consistent with his past class performance?
d. Am I familiar with past test instruments used to evaluate this child and how well can I rely on his prior test scores?
e. Have I observed this child in as many environments as possible (individual, large group, small group, play, home)?
f. Am I making illegitimate assumptions about this child? e.g., Do I assume he speaks and reads Spanish simply because he is Puerto Rican?
g. Have I actively sought additional information on non-school related variables that may have affected this child's school performance?
e.g., - health factors (adequate sleep, food)
- family difficulties
- peer group pressures

4. Does this child understand why he is in the assessment situation?
a. Have I tried to explain at his level of understanding what the reasons were for his referral?
b. Have I given this child the opportunity to freely express his perceptions of "the problem"?
c. Have I discussed with the child what next step activities may be involved?

SELECTION OF APPROACH FOR ASSESSMENT

1. Have I considered what the best assessment approach is for this child?
a. Considering the reasons for referral, do I need to utilize behavioral observations, interviews, informal techniques or standardized techniques or a combination of the above?
b. Have I given as much thought to assessing this child's adaptive behavior as I have to his academic school performance?
c. Are the approaches I am considering consistent with the child's receptive and expressive abilities?
d. Am I placing an overdependence on one technique and overlooking others that may be more appropriate?
e. Have I achieved a balance between formal and informal techniques in my selection?

2. If I have selected to use standardized instruments, have I considered all of the ramifications?
a. Am I testing this child simply because I've always used tests in my assessment procedure?

b. Am I administering a particular test simply because it is part of THE BATTERY?
c. Am I administering a test because I have been directed to do so by the Administration?
d. Does the instrument I've chosen include persons in the standardization sample from this child's cultural group?
e. Are subgroup scores reported in the manual?
f. Were there large enough numbers of this child's cultural group in the test sample for me to have any reliance on the norms?
g. Does the instrument I have selected assume a universal set of experiences for all children?
h. Does the instrument selected contain illustrations that are misleading and/or outdated?
i. Does the instrument selected employ vocabulary that is colloquial, regional and/or archaic?
j. Do I understand the theoretical basis of the instrument?
k. Will this instrument easily assist in delineating a recommended course of action to benefit this child?
l. Have I reviewed current literature regarding this instrument?
m. Have I reviewed research related to potential cultural influences on test results?

TEST ADMINISTRATION

1. Are there factors (attitude, physical conditions) which support the need to reschedule this child for evaluation at another time?

2. Could the physical environment of the test setting adversely affect this child's performance?
   - room temperature
   - noise
   - inadequate space
   - poor lighting
   - furnishings inappropriate for child's size

3. Am I familiar with the test manual and have I followed its directions?

4. Have I given this child clear directions?
   a. If his native language is not English, have I instructed him in his language?
   b. Am I sure that this child understands my directions?

5. Have I accurately recorded entire responses to test items, even though the child's answers may be incorrect, so that I might later consider them when interpreting his test scores?

6. Did I establish and maintain rapport with this child throughout the evaluation session?

SCORING AND INTERPRETATION

1. Have I examined each item missed by this child rather than merely looking at his total score?
   a. Is there a pattern to the types of items this child missed?
   b. Are the items missed free of cultural bias?
   c. If I omitted all items missed that are culturally biased, would this child have performed significantly better?

2. Am I aware that I must consider other factors in the interpretation of this child's scores?
   a. Have I considered the effect the child's attitude and/or physical condition may have had on his performance?
   b. Have I considered the effect that the child's lack of rapport with me may have had on his performance?
   c. Does my interpretation of this child's performance include observations?
   d. Do I realize that I should report and interpret scores within a range rather than as a number?

3. What confidence do I have in this child's test scores?
   a. Are test scores the most important aspect of this child's evaluation?
   b. Will I allow test scores to outweigh my professional judgment about this child?

CONSULTATION WITH TEAM MEMBERS AND OTHERS

1. Am I working as an integral member of a multidisciplinary team on behalf of this child?
   a. Have I met with the team to share my findings regarding this child?
   b. Are other team member's evaluation results in conflict with mine?
   c. Can I admit my discipline's limitations and seek assistance from other team members?
   d. Do I willingly share my competencies and knowledge with other team members for the benefit of this child?
   e. Has the team arrived at its conclusion as a result of team consensus or was our decision influenced by the personality and/or power of an individual team member?

2. Is the multidisciplinary team aware of its limitations?
   a. Are we aware of community resource personnel and agencies that might assist us in developing an educational plan for this child? Do we utilize such resources before, during, and after the evaluation?
   b. Do we on the team feel comfortable in including this child's parents in our discussions?

ASSESSMENT REPORT

1. Is my report clearly written and free of jargon so that it can be easily understood by this child, his parents, and teachers?

2. Does my report answer the questions asked in the referral?

3. Are the recommendations I have made realistic and practical for the child, school, teacher and parents?

4. Have I provided alternative recommendations?

5. Have I included in my report a description of any problems that I encountered and the effects of such during the assessment process?

INDIVIDUAL EDUCATIONAL PLAN

1. Are we making this child fit into an established program or are we developing an individualized educational plan appropriate for this child?
   a. Have we identified this child's strengths and weaknesses?

b. Have we specified long range goals and immediate objectives for this child?
c. Are we willing to assist the teacher in implementing this child's educational plan?
d. Have we stated when and how this child's progress will be evaluated and by whom?

FOLLOW UP

1. What are my responsibilities after we have written this child's educational plan?
   a. Have I discussed my findings and recommendations with this child's parents and explained their due process rights? Have I given the parents a written copy of this child's educational plan?
   b. Have I met with those working with this child to discuss the educational plan and to assist them in implementing its recommendations?
   c. Have I discussed my findings and recommendations with this child at his level of understanding?
   d. Can I help those working directly with the child to become more familiar with this child's social and cultural background?
   e. Have I sought this child's parents' permission for release of any confidential materials to other agencies and professionals?
   f. Will I periodically review this child's educational plan in regard to his actual progress so that any necessary changes can be made?

SOME FINAL THOUGHTS

1. Do I believe in the right to an appropriate education for all children?

2. Would I be comfortable if MY child had been involved in THIS assessment process?

3. Is there a willingness and desire on my part to actively participate in in-service activities that will lead to the further development of my personal and professional growth?

APPENDIX C

GUIDELINES FOR SPECIAL EDUCATION ASSESSMENT [1]

I. **Programming and Intervention in the Regular Classroom.**

A. *Basic Principle:* Prior to referral to special education diagnostic services solutions to classroom learning and adjustment problems should be attempted in the regular classroom.

B. *Basic Principle:* Various resource personnel, e.g., remedial reading specialists, curriculum consultants, counselors, psychologists, speech clinicians and social workers, should be available to assist teachers in developing educational procedures for meeting the child's needs in the regular classroom.

Considerations:

1. Are specially trained personnel available to assist classroom teachers and do these personnel provide assistance to teachers in developing alternative procedures in the regular classroom?
2. What changes are made in the regular classroom programs in order to serve children with diverse backgrounds and diverse characteristics?
3. What alternative materials and approaches, independent of special education, exist and have been attempted for children with learning and adjustment problems?
4. In cases referred to special education services, what evidence exists to confirm that attempts were made to solve the problem within the regular classroom? Were special personnel involved? Was an organized plan developed? Was the plan implemented? Was the plan given sufficient time to be successful?
5. Were efforts made to inform parents of the problem and attempted solutions, and were parents given an opportunity to contribute to solutions attempted in the regular classroom?

II. **Screening and Referral Phase.**

A. *Basic Principle:* Prior to formal diagnostic procedures, adequate information should be obtained which establishes the nature and extent of deviation from reasonable expectations.

Considerations:

1. Is the concern related to classroom learning or adjustment stated or restated specifically in behavioral terms rather than in terms of a special education category?
2. Is the concern related to current classroom learning or adjustment supported and illustrated by descriptive samples of behaviors?
3. Is consideration given to and evidence provided concerning the child's strengths within school and in other situations?
4. Are other sources of information considered systematically? Is this information consistent or inconsistent with the referral? Other sources of information should include the educational history (evaluations by previous teachers previous educational methods and materials used, previous grades), achievement test scores, previous evaluations by support personnel, previous and current social and emotional patterns of behavior, etc.
5. Do the above sources of information confirm the need for consideration of special education alternatives or does the information suggest that solutions should be attempted within the regular classroom?

---

[1] The Guidelines were developed by a committee appointed by George Garcia, Director, Urban Education Section of the Iowa Department of Public Instruction. Committee members included Daniel Reschly, Consultant, George Garcia, Jeff Grimes, Wilbur House, Marry Maitre, Pat O'Rourke, and Wayne Mooers. The Guidelines have not been approved officially by any division of the Iowa Department of Public Instruction.

B. *Basic Principle:* Parental involvement shall be obtained in all phases of referral, evaluation, and placement. Informed consent and due process procedures should be initiated early and followed throughout.
Considerations:
1. Are parents informed of the reasons for the referral in precise, meaningful langauge?
2. Have all communications been in the primary language of the home?
3. Does the school use a variety of means to solicit active parental participation in all phases of evaluation and staffing? Are parents informed of their rights to examine all relevant records?
4. Are parents provided with information concerning the activities and kind of decisions anticipated in evaluation and staffing along with estimates of time required, and specification of personnel responsible?

III. **Evaluation.**
A. *Basic Principle:* The evaluation of children referred for special education services should be conducted by a multidisciplinary team.
Considerations:
1. Is someone assigned the responsibility of coordinating the work of the team members including, a) evaluating the referral, b) determining the kind of information needed, c) assigning appropriately trained personnel to collect the data, d) facilitating communication among the team members?
2. Are interim procedures established for assisting the child and classroom teacher while the evaluation and staffing are conducted?

B. *Basic Principle: Multifactored Assessment.* Children should be assessed in all areas related to the suspected handicap including where appropriate health, vision, hearing, adaptive behavior, sociocultural background, emotional status, academic performance, aptitude (intelligence), language, and psychomotor ability. No single procedure such as IQ test results is used as the primary source of information, and the assessment procedures are used to identify areas of specific educational need. "Testing and evaluation materials and procedures used for the purposes of evaluation and placement of handicapped children must be selected and administered so as not to be racially or culturally discriminatory."
Considerations:
1. *Situational Assessment.* Is an assessment of the school or classroom environment conducted which includes a behavioral definition of the referral problem(s)? Are data collected on the frequency and magnitude of the problem(s), and a study made of the antecedent, situational, and consequent conditions related to the problem?
2. *Health History.* Are data collected on physical/health conditions which may be related to the learning problem? This information would include factors such as developmental history, disease and injury data, sensory data, sensory status, medication(s) used, and nutrition.
3. *Personal and Social Adjustment.* Is personal and social adjustment (adaptive behaviors) in the home, neighborhood, and broader community evaluated using formal and informal data collection procedures?
4. *Personal and Social Adjustment.* Is personal and social adjustment (adaptive behaviors) in the school setting evaluated with formal and informal data collection procedures?
5. *Primary Language.* Is the child's primary language dominance determined, and are the assessment procedures administered and interpreted in a manner consistent with the primary language data?
6. *Social and Cultural Background.* Is the sociocultural background of the child assessed systematically, and are the results of other assessment procedures interpreted in light of the sociocultural data?
7. *Educational Achievement: Norm-Referenced.* Is educational achievement assessed with norm-referenced instruments which yield valid information concerning the child's current performance in relation to grade level expectancies?
8. *Educational Achievement: Criterion-Referenced.* Is educational achievement assessed with criterion-referenced instruments or devices which provide valid information concerning specific skills and deficit areas?

9. *Aptitude.* Is academic aptitude, i.e., general intelligence, assessed with appropriate instruments available, consideration given to variations in performance over different factors of academic aptitude, and results interpreted in view of strengths and limitations of such measures?
10. *Psychoeducational Process.* Are psychoeducational processes and motor skills related to learning assessed, and the influence of these factors on the learning or adjustment problem considered (e.g., attention, eye-hand coordination, language, visual-motor, visual perception, auditory discrimination, etc.)?
11. *Other information.* Where appropriate, is information from other areas potentially important to placement and educational programming considered, e.g., career and vocational interests and aptitudes?

IV. **Staffing.**

A. *Basic Principles:* Placement decisions should be based upon information from a variety of sources (see previous section). Consideration of the information from the multifactored assessment should be documented in the staffing report. Placement decisions should be made by a group of persons including appropriate professional personnel and parents. The least restrictive alternative principle shall guide the selection of option for serving children.

Considerations:

1. What evidence exists which documents the consideration of a broad variety of information, including both strengths and deficits, in determing educational needs and selection of placement options?
2. Does the determination of educational needs and selection of placement option include the contributions of relevant professional personnel and parents?
3. Are current educational status and educational needs stated precisely and supported by data?
4. Are alternative options considered for meeting these needs including regular education with or without support services?
5. Are special education eligibility recommendations made in conformance with the criteria for primary handicapping condition as defined in the Department of Public Instruction Special Education Rules and Regulations?
6. In making the special education eligibility recommendations, did the multidisciplinary team consider a broad variety of information including adaptive behavior and sociocultural background? How did this information influence the recommendations concerning goals for intervention and placement option?
7. Are a variety of program options considered in view of the information from the multifactored assessment? For example, using information on adaptive behavior outside of school to choose between special classes and resource options for mild mental retardation.
8. What evidence supports the choice of program option as an appropriate alternative for meeting the child's needs?
9. Is an interim plan developed and implemented to assist the child in the regular classroom until the placement recommendations are carried out?
10. Do the special education personnel inform parents of the primary handicapping condition (if any) and explain the full range of available alternatives for meeting the child's needs?
11. Do parents contribute to decisions concerning the objectives of special education service selected?
12. Are there provisions for members of the multidisciplinary staffing team to express opinions which disagree with the decision of the majority? Are the dissenting opinions in written form expressing the reasons for disagreement?

APPENDIX D

ADAPTIVE BEHAVIOR INSTRUMENTS

1. American Association on Mental Deficiency Adaptive Behavior Scale for Children and Adults. Order from AAMD, 5101 Wisconsin Avenue, N.W., Washington, D.C., 20016.

2. Public School Version Adaptive Behavior Scale. Order from AAMD, 5101 Wisconsin Avenue, N.W., Washington, D.C., 20016.

3. Camelot Behavioral Checklist. Order from Camelot Behavioral Systems, P.O. Box 3447, Lawrence, KS, 66044. Also Erdmark Associates, 1329 Northup Way, Bellevue, WA, 98005.

4. Cain-Levine Social Competency Scale. Order from Consulting Psychologists Press, 577 College Avenue, Palo Alto, CA, 94306.

5. Balthazar Scales of Adaptive Behavior Consulting Psychologists Press; see above.

6. Adaptive Behavior Inventory for Children (Part of SOMPA) Psychological Corporation, 757 Third Avenue, New York, NY, 10017.

7. Vineland Social Maturity Scale. Order from American Guidance Services, Publishers Building, Circle Pines, MN, 55014.

8. Children's Adaptive Behavior Scale. Order from Humanics, Ltd., P.O. Box 7447, Atlanta, GA, 30309.

REFERENCES

APA Monitor. Exhibit A: IQ Trial. Plaintiffs Take the Stand. December, 1977, p. 4.

Anastasi, A. Psychological testing (4th ed.). New York: MacMillan, 1976.

Ashurst, D., & Meyers, E. Social system and clinical model in school identification of the educable retarded. In Eyman, R., Meyers, E., and Tarjan, G. (Eds.), Sociobehavioral studies in mental retardation. Washington, D.C.: American Association on Mental Deficiency, 1973.

Baller, W., Charles, D., & Miller, E. Mid-life attainment of the mentally retarded. Genetic Psychology Monographs, 1967, 75, 235-329.

Bartel, N., Grill, J., & Bryen, D. Language characteristics of black children: Implications for assessment. Journal of School Psychology, 1973, 11, 351-364.

Benton, A. The concept of pseudofeeblemindedness. Archives of Neurology and Psychiatry, 1956, 75, 379-388.

Bergan, J. Behavioral Consultation. Columbus, Ohio: Charles E. Merrill, 1977.

Bernal, E. A response to "Educational uses of tests with disadvantaged subjects". American Psychologist, 1975, 30, 93-95.

Bersoff, D. Regarding psychologists testily: Legal regulation of psychological assessment in the public schools. Maryland Law Review, 1979, 39, 27-120.

Brody, E., & Brody, N. Intelligence: Nature, determinants, and consequences. New York: Academic Press, 1976.

Brown, F. The algebra works - But what does it mean? School Psychology Digest, 1979, 8, 213-218.

Buckley, K., & Oakland, T. Contrasting localized norms for Mexican-American children on the ABIC. Paper presented at the annual meeting of the American Psychologcal Association, San Francisco, August, 1977.

Clarizio, H. In defense of the IQ test. School Psychology Digest, 1979, 8, 79-88. (a)

Clarizio, H. Commentary on Mercer's rejoinder to Clarizio. School Psychology Digest, 1979, 8 207-209.

Cleary, T. A., Humphreys, L., Kendrick, A., & Wesman, A. Educational uses of tests with disadvantaged students. American Psychologist, 1975, 30, 15-41.

Cone, J., & Hawkins, R. (Eds.), Behavioral assessment: New directions in clinical psychology. New York: Brunner/Mazel, 1977.

CORRC (Coordinating Office on Regional Resource Centers). With bias toward none: Volune III, A national survey of assessment programs and procedures. Lexington: University of Kentucky, Undated (distributed in 1979).

Coulter, A., & Morrow, H. The concept and measurement of adaptive behavior. New York: Grune & Stratton, 1978.

Cromwell, R., Blashfield, R., & Strauss, J. Criteria for classification systems. In N. Hobbs (Ed.), Issues in the classification of children. San Fransciso: Jossey-Bass, 1975.

Cronbach, L. Essentials of psychological testing (3rd ed.). New York: Harper & Row, 1970.

Cronbach, L. Five decades of public controversy over mental testing. American Psychologist, 1975, 30, 1-14.

Dean, R. S. Reliability of the WISC-R with Mexican-American children. Journal of School Psychology, 1977, 15, 267-268.

Deer, P. Fry Bread IQ Test: American Indian Intelligence Test. Camp Crier, 1980 (April), pp. 8-10.

Dent, H. Assessing black children for mainstream placement. In R. Jones (Ed.), Mainstreaming and the minority child. Reston, VA: Council for Exceptional Children, 1976.

Deutsch, M., Fishman, J., Kogan, I., North, R., & Whitman, M. Guidelines for testing minority group children. Journal of Social Issues, 1964, 20, 129-145.

Diana vs. State Board of Education, C-70 37 RFP, District Court for Northern California (February, 1970).

Doll, E. Measurement of social competence. Circle Pines, MN: American Guidance Service, 1953.

Dove, O. O. Dove Counterbalanced Intelligence Test. Mimeo, n.d.

Dunn, L. Special education for the mildly retarded: Is much of it justifiable? Exceptional Children, 1968, 35, 5-22.

Farber, B. Mental retardation: Its social context and social consequences. Boston: Houghton-Mifflin, 1968.

Federal Register, August 23, 1977, p. 42474-42518. Education of handicapped children. Regulations Implementing Education for all Handicapped Children Act of 1975 (Public Law 94-142).

Fisher, A. Four approaches to classification of mental retardation. Paper presented at the annual meeting of the American Psychological Association, August, 1978.

Fisher, A. Administrative interventions to provide educational services for declassified students. Paper presented at the annual meeting of the American Psychological Association, New York, August, 1979.

Flaugher, R. The many definitions of test bias. American Psychologist, 1978, 33, 671-679.

Gallagher, J. The special education contract for mildly handicapped children. Exceptional Children, 1972, 38, 527-535.

Garber, H. Intervention in infancy: A developmental approach. In M. Begab & S. Richardson (Eds.), The mentally retarded and society: A social science perspective. Baltimore: University Park Press, 1975.

Goldman, R., & Hartig, L. The WISC may not be a valid predictor of school performance for primary-grade minority children. American Journal of Mental Deficiency, 1976, 80, 583-587.

Goodman, J. Is tissue the issue? A critique of SOMPA's models and tests. School Psychology Digest, 1979, 8, 47-62.

Gridley, G., & Mastenbrook, J. Research on the need for local norms for the Adaptive Behavior Inventory for Children. Paper presented at the annual meeting of the American Psychological Association, San Francisco, August, 1977.

Grossman, H. (Ed.), Manual on terminology and classification in mental retardation. Washington, D.C.: American Association on Mental Deficiency, 1973; 1977.

Guadalupe v. Tempe Elementary School District, 71-435, District Court for Arizona, January, 1972.

Hartlage, L., & Steele, C. WISC and WISC-R correlates of academic achievement. Psychology in the Schools, 1977, 14, 15-18.

Havighurst, R. Using the IQ wisely. NEA Journal, 1951, 15, 540-541. (Quote and reference taken from Platt & Bardon, 1967).

Heber, R. Mental retardation: Concept and classification. In E. Trapp & P. Himelstein, Readings on the exceptional child. New York: Appleton-Century-Crofts, 1962.

Heber, R. A manual on terminology and classification in mental retardation (2nd ed.). American Journal of Mental Deficiency, 1961, Monograph Supplement 64.

Hunt, J. Intelligence and experience. New York: Ronald, 1961.

Hunter, J., & Schmidt, F. Critical analysis of the statistical and ethical implications of various definitions of test bias. Psychological Bulletin, 1976, 83, 1053-1071.

Hynd, G., & Garcia, W. Intellectual assessment of the Native American student. School Psychology Digest, 1979, 8, 446-454.

Jackson, G. On the report of the ad hoc committee on educational uses of tests with disadvantaged students. American Psychologist, 1975, 30, 88-92.

Jensen, A. How much can we boost IQ and scholastic achievement? Harvard Education Review, 1969, 39, 1-123.

Jensen, A. Educability and group differences. New York: Harper & Row, 1973.

Jones, R. Protection in evaluation procedures: Criteria and recommendations. In L. Morra (Ed.), Developing criteria for evaluation of the protection in evaluation procedures provision of Public Law 94-142. Washington, D.C.: U.S. Office of Education, Bureau of Education for the Handicapped, 1978.

Kamin, L. The science and politics of IQ. Potomac, MD: Erlbaum Associates, 1974.

Kaufman, A. Factor analysis of the WISC-R at 11 age levels between 6½ and 16½ years. Journal of Consulting and Clinical Psychology, 1975, 43, 135-147.

Kaufman, A. Intelligent testing with the WISC-R. New York: Wiley-Interscience, 1979. (a)

Kaufman, A. WISC-R research: Implications for interpretation. School Psychology Digest, 1979, 8, 5-27. (b)

Kaufman, A., & Doppelt, J. Analysis of WISC-R standardization data in terms of stratification variables. Child Development, 1976, 47, 165-171.

Kazimour, K., & Reschly, D. Investigation of norms and concurrent validity for the Adaptive Behavior Inventory for Children. American Journal of Mental Deficiency, in press.

Keller, H. Issues in the use of observational assessment. School Psychology Review, 1980, 9, 21-30.

Kolstoe, O. Teaching educable mentally retarded children (2nd ed.). New York: Holt, 1976.

Koppitz, E. The Bender Gestalt Test for Young Children, Volume II: Research and Application, 1963-1973. New York: Grune & Stratton, 1975.

Lambert, N. Contributions of school classification, sex, and ethnic status to adaptive behavior assessment. Journal of School Psychology, 1979, 17, 3-16.

Lambert, N., Windmiller, M., Cole, L., & Figueroa,, R. Manual for the public school version of the AAMD adaptive behavior scale. Washington, D.C.: American Association on Mental Deficiency, 1975.

Larry P. et. al. vs. Wilson Riles et. al. United States District Court, Northern District of California, Case No. C-71-2270 RFPP, 1972, 1974, 1979.

Leland, H. Theoretical considerations of adaptivee behavior. In A. Coulter & H. Morrow (Eds.), The concept and measurement of adaptive behavior. New York: Grune & Stratton, 1978.

Leland, H., Shellhaas, M., Nihira, K., & Foster, RR. Adaptive behavior. A new dimension in the classification of the mentally retarded. Mental Retardation Abstracts, 1967, 4, 359-387.

Lesser, G., Fifer, G., & Clark, D. Mental abilitiies of children from different social-class and cultural groups. Monographs of the Society for Research in Child Development, 1965, 30(4), 1-115.

Loehlin, J., Lindzey, G., & Spuhler, J. Race diffferences in intelligence. San Francisco: Freeman, 1975.

MacMillan, D. Mental retardation in school and scociety. Boston: Little, Brown, & Co., 1977.

MacMillan, D., Jones, R., & Aloia, G. The mentally retarded label: A theoretical analysis and review of research. American Journal of Mental Deficiency, 1974, 79, 241-261.

Matarazzo, D. Wechsler's measurement and appraisal of adult intelligence: 5th and enlarged edition. Baltimore: Williams and Wilkins, 1972.

Matluck, J., & Mace, B. Language characteristics of Mexican-American children: Implications for assessment. Journal of School Psychology, 1973, 11, 365-386.

Mattie, T., et. al. vs. Charles E. Holladay, et. al. C.A. No. DC-75-31-S. United States District Court for the Northern District of Mississippi, 1979.

McCall, R., Appelbaum, M., & Hogarty, P. Developmental changes in mental performance. Monographs of the Society for Research in Child Development, 1973, 38, Whole No. 150.

McCandless, B. Children: Behavior and development (2nd ed.). New York: Holt, 1967.

Meichenbaum, D. Cognitive behavior modification. New York: Plenum, 1977.

Mercer, J. Sociological perspectives on mild mental retardation. In H. Haywood (Ed.), Socio-cultural aspects of mental retardation. New York: Appleton-Century-Crofts, 1970.

Mercer, J. Labelling the mentally retarded. Berkeley, CA: University of California Press, 1973.

Mercer, J. Protection in evaluation procedures. In Morra, L. (Ed.), Developing criteria for the evaluation of protection in evaluation procedures provisions of Public Law 94-142. Washington, D.C.: U. S. Office of Education, Bureau of Education for the Handicapped, 1978.

Mercer, J. Technical manual: SOMPA: System of multicultural pluralistic assessment. New York: Psychological Corporation, 1979.

Meyers, C., MacMillan, D., & Yoshida, R. Validity of psychologists identification of EMR students in the perspective of the California decertification experience. Journal of School Psychology, 1978, 16, 3-15.

Meyers, C., Sundstrom, P., & Yoshida, R. The school psychologist and assessment in special education: A report of the Ad Hoc Committee of APA Division 16. Monographs of Division 16 of the American Psychological Association, 1974, 2(1), 3-57.

Nelson, R. The use of intelligence tests within behavioral assessment. Behavioral Assessment, 1980, 2, 417-423.

Oakland T. (Ed.), Psychological and educational assessment of minority children. New York: Brunner/Mazel, 1977. (a)

Oakland, T. Pluralistic norms and estimated learning potential. Paper presented at the annual meeting of the American Psychological Association, San Francisco, 1977. (b)

Oakland, T. An evaluation of the ABIC pluralistic norms, and Estimated Learning Potential. Journal of School Psychology, 1980, 18, 3-11.

Oakland, T., & Goldwater, D. Assessment and interventions for mildly retarded and learning disabled children. In G. Phye & D. Reschly (Eds.), School Psychology: Perspectives and Issues. New York: Academic Press, 1979.

(PASE) Parents in Action on Special Education vs. Joseph P. Hannon. U. S. District Court, Northern District of Illinois, Eastern Division, No. 74 (3586), July, 1980.

Patrick, J., & Reschly, D. Relationship of state education criteria and demographic variables to prevalence of mental retardation. Unpublished manuscript, Department of Psychology, Iowa State University, 1980.

(PCMR) President's Committee on Mental Retardation. The six-hour retarded child. Washington, D.C.: United States Government Printing Office, 1970.

Petersen, N., & Novick, M. An evaluation of some models for culture fair selection. Journal of Educational Measurement, 1976, 13, 3-29.

Platt, H., & Bardon, J. The school psychologist's role with the mentally retarded. In J. Magary (Ed.), School psychological services in theory and practice: A Handbook. Englewood Cliffs, N.J.: Prentice-Hall, 1967.

Rapaport, D., Gill, M., & Schafer, R. Diagnostic psychological testing (Revised edition by R. Holt). New York: International Universities Press, 1968.

Reschly, D. Comparisons of bias in assessment with conventional and pluralistic measures. Paper presented at the Council for Exceptional Children Annual Convention, 1978. Also available from ERIC Document Reproduction Service, ED 153-386.

Reschly, D. WISC-R factor structures among Anglos, Blacks, Chicanos, and Native American Papagos. Journal of Consulting and Clinical Psychology, 1978, 46, 417-422.

Reschly, D. Nonbiased assessment. In G. Phye & D. Reschly (Eds.), School psychology: Perspectives and issues. New York: Academic Press, 1979.

Reschly, D. Psychological evidence in the Larry P. opinion: A case of right problem-wrong solution? School Psychology Review, 1980, 9, 136-148.

Reschly, D. Concepts of bias in assessment and WISC-R research with minorities. In A. Engin (Ed.), WISC-R handbook. Washington, D.C.: National Association of School Psychologists, 1980.

Reschly, D. Nonbiased assessment and the mildly retarded. In T. Oakland, Nonbiased assessment. National School Psychology Inservice Training Network, University of Minnesota - Federal Bureau of Education for the Handicapped, 1980.

Reschly, D. Sociocultural background, adaptive behavior, and concepts of bias in assessment. In C. Reynolds & T. Gutkin (Eds.), Handbook of school psychology. New York: Wiley-Interscience, 1981.

Reschly, D. Evaluation of the effects of SOMPA measures on classification of students as mildly retarded. American Journal of Mental Deficiency, In press.

Reschly, D., & Reschly, J. Validity of WISC-R factor scores in predicting teacher ratings of achievement and attention among four groups. Journal of School Psychology, 1979, 17, 355-361.

Reschly, D., & Rosa-Reynolds, J. Report: Iowa Assessment Project. Unpublished manuscript , Department of Psychology, Iowa State University, 1980.

Reynolds, C., & Gutkin, T. Predicting the premorbid intellectual status of children using demographic data. Clinical Neuropsychology, 1979, 1, 36-38.

Richmond, B., & Kicklighter, R. Children's Adaptive Behavior Scale. Atlanta: Humanics Limited, 1980.

Robinson, N., & Robinson, H. The mentally retarded child (2nd ed.). New York: McGraw-Hill, 1976.

Salvia, J., & Ysseldyke, J. Assessment in special and remedial education. Boston: Houghton-Mifflin, 1978.

Samuda, R. Psychological testing of American minorities: Issues and consequences. New York: Dodd, Mead and Co., 1975.

Sandoval, J. The WISC-R and internal evidence of test bias with minority groups. Journal of Consulting and Clinical Psychology, 1979, 47, 919-927.

Sandoval, J., & Mille, M. Accuracy of judgments of WISC-R item difficulty for minority groups. Journal of Consulting and Clinical Psychology, 1980, 48, 249-253.

Sattler, J. Racial "experimenter effects" in experimentation, testing, interviewing, and psychotherapy. Psychological Bulletin, 1970, 73, 137-160.

Sattler, J. Intelligence testing of ethnic minority-group and culturally disadvantaged children. In L. Mann & D. Sabatino (Eds.), The first review of special education, Vol. 2. Philadelphia: The JSE Press, 1973.

Sattler, J. Assessment of children's intelligence. Philadelphia: S. B. Saunders, 1974.

School Psychology Digest, 1979, Volume 8, No. 1 & 2. SOMPA: A Symposium.

School Psychology Review, 1980, Volume 9, No. 2, Larry P. vs. Riles Case.

Scott, L. Identification of declassified students: Characteristics and needs of the population. Paper presented at the annual meeting of the American Psychological Association, New York, August, 1979.

Shockley, W. Models, mathematics, and the moral obligation to diagnose the origin of Negroe IQ deficits. Review of Educational Research, 1971, 41, 369-377.

Silverstein, A. Alternative factor analytic solutions for the Wechsler Intelligence Scale for Children - Revised. Educational and Psychological Measurement, 1977, 37, 121-124.

Talley, R. Evaluating the effects of implementing SOMPA. Bloomington, Indiana: School of Education, University of Indiana, Center for Innovation in Teaching the Handicapped, November, 1978.

Thorndike, R. Concepts of culture-fairness. Journal of Educational Measurement, 1971, 8, 63-70.

Tomlinson, J., Acker, N., Canter, A., & Lindborg, S. Minority status, sex, and school psychological services. Psychology in the Schools, 1977, 14, 456-460.

Trotman, F. Race, IQ, and the middle class. Journal of Educational Psychology, 1977, 69, 266-273.

Tucker, J. Operationalizing the diagnostic-intervention process. In T. Oakland (Ed.), Psychological and educational assessment of minority children. New York: Brunner/Mazel, 1977.

Turnbull, H. The past and future impact of court decisions in special education. Kappan, 1978, 60, 523-527.

Williams, R. Abuses and misuses in testing black children. The Counseling Psychologist, 1971, 2, 62-73.

Williams, R. The BITCH-100: A culture specific test. Journal of Afro-American Issues, 1975, 3, 103-116.

Yoshida, R., MacMillan, D., & Meyers, C. The decertification of minority group EMR students in California: Student achievement and adjustment. In R. Jones (Ed.), Mainstreaming and the minority child. Reston, VA: Council for Exceptional Children, 1976.

Ysseldyke, J. Issues in psychoeducational assessment. In G. Phye & D. Reschly (Eds.), School psychology: Perspectives and issues. New York: Academic Press, 1979.

Zigler, E. Familial mental retardation. A continuing dilemma. Science, 1967, 155, 292-298.

Ziskin, J. Coping with psychiatric and psychological testimony. Beverly Hills: Law and Psychology Press, 1975.